TEKNILLINEN KORKEAKOULU
Tietotekniikan osasto
Informaatiotekniikan laboratorio

Lasse Rasinen

Aikakauslehtien myyntipistekohtainen valikoimanoptimointi

Diplomi-insinöörin tutkintoa varten tarkastettavaksi jätetty diplomityö.

Espoo, 2.12.2004

Työn valvoja: Ma. prof. Jaakko Hollmén
Työn ohjaaja: DI Kimmo Kiviluoto

TEKNILLINEN KORKEAKOULU
Tietotekniikan osasto

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Lasse Rasinen
Työn nimi:	Aikakauslehtien myyntipistekohtainen valikoimanoptimointi
Päivämäärä:	2.12.2004 **Sivuja:** 59
Professuuri:	T-122 Informaatiotekniikka
Työn valvoja:	Ma. prof. Jaakko Hollmén
Työn ohjaaja:	DI Kimmo Kiviluoto

Optimaalisen tuotevalikoiman määrittäminen on perinteisesti ollut hyvin kokemuspohjainen ala. Pelkästään yhden tuotteen kysynnän ennustaminen on ollut pitkään tunnettu ongelma, johon kirjallisuudesta löytyy useita ratkaisuja. Nämä ratkaisut eivät kuitenkaan ota huomioon tuotteiden välisiä vaikutussuhteita. Tässä diplomityössä etsittiin usean tuotteen tapauksessa toimivia strategioita tuotevalikoiman määrittämiseen aikakauslehtien irtonumeromyynnissä.

Työ tehtiin Xtract Oy:ssä Lehtipisteen asiakasprojektina, ja Lehtipiste toimitti suurimman osan käytettävästä data-aineistosta. Rajausten jälkeen käytettävissä olevaan aineistoon kuului 1498 aikakauslehteä, 5009 myyntipistettä ja näihin liittyvä myyntitapahtuma-aineisto kahden ja puolen vuoden ajalta.

Työssä tutkitaan ensiksi olemassaolevan kirjallisuuden ratkaisuehdotuksia ja aikakauslehtimarkkinoihin liittyviä erikoisolosuhteita. Toiseksi työssä analysoidaan saatavilla olevaa dataa ja pyritään löytämään ennustamisen kannalta olennaisia tekijöitä.

Lopuksi työssä kehitetään menetelmä, jolla voidaan laatia myyntipisteelle ehdotus uudeksi valikoimaksi samalla tavalla käyttäytyvien myyntipisteiden avulla. Menetelmää testattiin kolmen kuukauden ajan rajatussa joukossa myyntipisteitä. Testin tuloksena saatiin 2,36 % parannus myyntiin. Kehitetty menetelmä on tietyin varauksin sovellettavissa myös muihin toimialoihin.

Työssä käytettiin itseorganisoivaa karttaa datan visualisointiin, lineaarisia ja yleistettyjä lineaarisia regressiomalleja ennustamiseen ja k:n lähimmän naapurin menetelmää lopullisten tulosten määrittämiseen.

Avainsanat: aikakauslehdet, valikoimanoptimointi, itseorganisoiva kartta, k:n lähimmän naapurin menetelmä.

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Computer Science and Engineering

ABSTRACT OF THE MASTER'S THESIS

Author: Lasse Rasinen

Name of the thesis: Magazine Sales Outlet Assortment Optimization

Date: December 2, 2004 **Number of pages:** 59

Professorship: T-122 Computer and Information Science

Supervisor: Professor (pro tem) Jaakko Hollmén

Instructor: Kimmo Kiviluoto, M.Sc. (Tech)

Determining the optimal assortment in retail has traditionally required skill and experience in the field. Forecasting the demand for a single product is a well-known problem, and several solutions are described in the literature. However, these solutions disregard interproduct relations. This thesis explores possible strategies for determining the optimal assortment for several magazines in Finnish single copy market.

The thesis was done in Xtract Ltd as a customer project for Lehtipiste; Lehtipiste also supplied most of the data used in the project. The data consists of 1498 magazines, 5009 sales locations, and related sales event data for a period of two and half years.

The thesis first examines the solutions presented in earlier works and the special conditions associated with magazine sales. The thesis then analyses the available data and seeks the important elements for forecasting.

Finally the thesis shows a method for developing a suggestion for a new assortment for a sales location by examining sales locations with similar behaviour. The method is tested for three months in a selected group of sales locations, with 2.36 % improvement in sales. The method is also applicable in other industries with some caveats.

The thesis used the Self-Organizing Map for visualising the data, linear and generalized linear regression models for forecasting and the k nearest neighbour method for producing the final results.

Keywords: magazines, assortment optimization, self-organizing map, k nearest neighbour method

Kiitokset

Kiitokset Lehtipisteelle mahdollisuudesta tehdä näin mielenkiintoista projektia; haluaisin kiittää kaikkia Lehtipisteen henkilökunnasta projektiin osallistuneita, erityisesti Jaakko Vuorista ja Arja Happosta.

Haluan myös kiittää Xtract Oy:n porukkaa mainiosta työympäristöstä ja motivoinnista, erityisesti Juusoa ja Kimmoa sparrauksesta ja kannustustuksesta tämän diplomityön loppuun saattamisessa.

Oikolukukiitokset: tekn. yo A. Rasinen, tekn. yo. H. Pykälä.

Ilman Herttoniemen Subwayn ravitsevia leipiä kirjoitustyö olisi huomattavasti hidastunut; siispä kiitokset heillekin.

Siviilielämän puolella kiitokset FH Strike Teamille saunailloista ja kannustuksesta: Tuuli, Antti, Antti, Jaakko, Heikki ja Nikolaj (ja myös Jan, Salomon, Antti ja Jere).

Ja erityisesti kiitos Helille pitkästä pinnasta ja kärsivällisyydestä diplomityön loppurutistuksen aikaan.

Otaniemi, 2.12.2004

Lasse Rasinen

Sisältö

Taulukot

Kuvat

Lyhenteet

CAM	Cluster Abstraction Model (HACM:in uusi nimi)
DFU	Demand Forecasting Unit
EM	Expectation-Maximization (-algoritmi)
GMROI	Gross-Margin Return On Investment
HACM	Hierarchical Asymmetric Clustering Model
LDA	Latent Dirichlet Allocation
LSA/LSI	Latent Semantic Analysis/Indexing
MAD	Mean Absolute Deviation
mPCA	multinomial-PCA
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PLSA/PLSI	Probabilistic Latent Semantic Analysis/Indexing
SKU	Stock-keeping Unit
SOM	Self-Organizing Map

Symbolit ja merkinnät

a	Aiheryhmä
$a(l)$	Lehden l aiheryhmä
m	Myyntipiste
l	Lehti
N_l	Lehden l ilmestymiskerrat vuodessa
M	Kaikkien myyntipisteiden joukko
L	Kaikkien lehtien joukko
$n_m(l)$	Lehden l numerokohtainen myynti kappalemääräisenä myyntipisteessä m
$v_m(l)$	Lehden l keskimääräinen viikkomyynti myyntipisteessä m
$v_m(a)$	Aiheryhmään a kuuluvien lehtien yhteenlaskettu euromääräinen viikkomyynti myyntipisteessä m
$p_m(a)$	Aiheryhmän a suhteellinen osuus myyntipisteessä m
$g_X(l)$	Erilaiset hyvyysarviot lehdelle l; X:n merkitys selitetään tapauskohtaisesti
$g_m(l)$	Lehden l hyvyysarvio myyntipisteessä m

Luku 1

Johdanto

> Kaupassa oli hyllykaupalla ihanimman näköisiä makeisia mitä kuvitella saattaa. Kermaisia nougatlohkareita, hohtavan pinkkejä kookosjääkuutioita, paksuja hunajanvärisiä toffeekimpaleita, satamäärin suklaapatukoita siisteissä riveissä; iso tynnyrillinen joka maun rakeita ja toinen viuhpiiporeita, Ronin mainitsemia leijuttavia mehujauhepalloja.
> — J. K. Rowling, *Harry Potter ja Azkabanin vanki*

Tämä työ tehtiin Xtract Oy:ssä Lehtipisteen tilauksesta. Työn puitteissa tutkittiin erilaisten analyysimenetelmien, kuten SOM-kartan tai k-NN -menetelmän käyttöä myyntipisteiden aikakauslehtivalikoimien ennustamiseen. Tavoitteena oli analysoida saatavilla olevaa dataa eri lehtien ja myyntipisteiden myyntikäytöksestä, ja kehittää tämän analyysin pohjalta menetelmä parempien valikoimien ennustamiseen.

1.1 Lehtipisteen taustaa

Lehtipiste toimii Suomessa monopoliasemassa aikakaus- ja iltapäivälehtien jakelijana kustantajilta myyntipisteisiin ympäri Suomea. Lehtipiste toimii kustantajilta saamiensa komissiopalkkioden varassa kustantajien ottaessa taloudellisen vastuun myymättä jääneistä lehdistä. Jakelukustannukset ovat puolestaan Lehtipisteen vastuulla.

Aikakauslehti on luonteeltaan *häviävä* tuote; lehden tiettyä numeroa ei ole saatavilla ennen sen julkaisua, ja rajallisen myyntiajan jälkeen numero otetaan pois myynnistä ja korvataan uudella numerolla. Vastaavanlaisia markkinoita ovat muun muassa erilaiset sesonkituotteet, esimerkiksi joulutuotteet, tuoreet elintarvikkeet tai matkaliput.

Aikakauslehtimarkkinoilla on kuitenkin muutama huomionarvoinen erikoisolosuhde: koska Lehtipisteellä ja yksittäisillä myyntipisteillä ei ole huomattavaa taloudellista sitoumusta lehtiin, lehtiä ei myyntiajan vähentyessä myydä alennushintaan varastojen tyhjentämiseksi. Lehtipiste myös tyypillisesti toimittaa lehdet myyntipisteisiin vain uuden numeron ilmestyessä; täydennystoimituksia ei yleensä tehdä.

Myös Suomen kaksikieliset alueet ovat ongelmallisia, koska on hyvin yleistä, että paikkakunnalla suomenkielinen ja ruotsinkielinen väestö suosivat omia kauppojaan. Tällöin demografiatiedoista ei ole suurta hyötyä, vaan toista kauppaa on kohdeltava pääosin suomenkielisenä ja toista pääosin ruotsinkielisenä.

1.2 Tavoitteet

Projektin alussa Lehtipisteen toiveena oli saada automatisoitu analytiikkaratkaisu Lehtipisteen valikoimanhallinnan tueksi. Lehtipisteen nykyinen myyntipisteiden valikoimanhallinta luottaa hyvin pitkälle yksittäisten piiriedustajien kokemukseen ja osittain myyntipisteiden omaan aktiivisuuteen epäkohtien korjaamisessa. Tavoitteena oli myynnin kasvattaminen huonojen valikoimien parantamisen kautta ja kustannussäästöjen saavuttaminen ihmisvoimin tehtävän työmäärän vähentyessä.

Projektin alussa Lehtipiste esitti seuraavanlaisia yksilöityjä ominaisuustoiveita. Mallin tulisi

- optimoida valikoima yksittäisessä myyntipisteessä siten, että kokonaismyynti maksimoituu,
- simuloida myyntiä jossain myyntipisteessä eri lehtivalikoimilla,
- tunnistaa myyntipisteet, joiden valikoima ei ole optimaalinen; samalla voidaan raportoida ero nykymyynnin ja optimaalisen myynnin välillä, ja
- löytää optimaaliset myyntipisteet jollekin uudelle lehdelle.

1.3 Toteutus

Näiden toiveiden perusteella projektille valittiin tavoitteeksi kehittää malli, joka ennustaisi käytettävissä olevan datan pohjalta myyntivalikoimaan tehtävien muutosten vaikutukset. Mallia testattaisiin ensin kylmäharjoitteluna

myyntipiste- ja lehtikohtaisten listojen sekä historiadatasta ennustettujen tulosten nojalla. Tämän jälkeen suoritettaisiin käytännön testi, jossa tutkittaisiin suositusten toimivuutta ja käyttökelpoisuutta 2–3 myyntipiirissä (n. 250 myyntipistettä) 3 kk testijaksolla.

Työnjako projektissa Lehtipisteen ja Xtract:in kesken oli seuraava: Lehtipiste toimitti tarvittavan tietoaineiston, jota Xtract rikasti omalla Xtract Consumer LifeCycles -aineistollaan. Varsinainen mallinnustyö oli Xtract:in vastuulla, ja Lehtipiste ohjasi mallinnusta antamalla kommentteja ja määrittämällä mallin toimintaan vaikuttavia nykyisin käytössä olevia liiketoimintasääntöjä.

Tulokset testattiin valitsemalla käsin joukko myyntipisteitä, joiden valikoiman optimointiin mallia käytettiin. Lehtipisteelle toimitettiin lista näiden myyntipisteiden valikoimiin tehtävistä poistoista ja lisäyksistä. Samanaikaisesti valittiin joukko samankaltaisia verrokkimyyntipisteitä, joiden valikoimaa ei optimoitu. Mallin toimivuus selvitettiin vertaamalla myynnin kehitystä näissä kahdessa eri joukossa.

1.4 Rakenne

Tässä työssä tutkitaan ensin jo tehtyjä tutkimuksia vähittäismyynnin alalta yleisesti sekä nimenomaan Lehtipisteelle tehtyjä töitä. Lisäksi perehdytään mahdollisesti sovelluskelpoisiin mallinnustekniikoihin. Kappaleessa 3 tutkitaan tärkeimpiä tässä sovelluksessa käytettyjä matemaattisia menetelmiä yleisellä tasolla.

Kappaleessa 4 käsitellään saatavilla olevaa aineistoa sekä sivutaan tehtyjä esikäsittelyoperaatiota. Kappaleessa 5 tutkitaan eri muuttujatyyppien ja menetelmien soveltuvuutta käsillä olevaan ongelmaan ja selostetaan lopullinen malli syötemuuttujineen ja optimointikriteereineen.

Kappaleessa 6 kuvataan mallille tehtyjä perustoimivuustestejä, selostetaan kenttäkoejärjestelyt ja esitetään kenttäkokeiden tulokset. Lopuksi kappaleessa 7 käydään läpi mahdollisia parannuskohteita sekä esitetään yhteenveto työstä.

Luku 2

Kirjallisuuskatsaus

Tämä kirjallisuuskatsaus jakautuu kolmeen osaan: ensimmäisessä osassa perehdytään vähittäiskauppaan yleisesti ja kyseisellä toimialalla hyväksi havaittuihin toimintatapoihin ja ennustusmenetelmiin. Toisessa osassa tutkitaan puolestaan Lehtipisteen toimeksiannosta aiemmin tehtyjä tutkimuksia mahdollisten yhtymäkohtien löytämiseksi. Lopuksi käsitellään ongelman kannalta relevantteja nykyisiä tutkimuskohteita ja matemaattisia menetelmiä.

2.1 Vähittäiskauppa toimialana

Seuraava osuus perustuu pääosin Michael Levyn ja Barton Weitzin kirjaan *Retail Management* [25]. Kyseinen teos käsittelee vähittäistavarakaupan toimialaa pääasiassa tavaratalojen ja supermarkettien näkökulmasta. Kirjassa on myös havaittavissa painotusta vaatekaupan suuntaan. Tämän diplomityön puitteissa kirjan oleellisin osuus on materiaalinhallinta, erityisesti *valikoimansuunnittelu* (*assortment planning*).

2.1.1 Valikoimanhallinnan peruskäsitteitä

Valikoimanhallinnassa tuotteet on tyypillisesti jaettu hierarkiaan, jolloin ostojen suunnittelu ja vastuukysymykset voidaan järkevämmin hallita. Kirja esittelee esimerkkinä amerikkalaisen National Retail Federationin käyttämän viisiportaisen hierarkian:

- *Merchandise group* (tavararyhmä)
- *Department* (osasto)

- *Classification* (luokittelu)
- *Category* (kategoria)
- *Stock-keeping unit, SKU* (tuote)

Hierarkian kaksi ylintä tasoa ovat sidoksissa yritysrakenteeseen, joten niitä ei tässä käsitellä sen tarkemmin. *Luokitteluun* lasketaan samalla tavalla käyttäytyvät tuotteet; esimerkiksi housut ja takit muodostavat kumpikin oman luokittelunsa. *Kategoriaan* lasketaan sellaiset tuotteet, joiden voidaan katsoa olevan toistensa korvikkeita (farkut, puvun housut).

Tuotevalikoiman suunnittelussa lähdetään liikkelle asetetuista taloudellisista tavoitteista. Kirja suosittelee käytettäväksi taloudelliseksi mittariksi *sijoitetun pääoman bruttorajatuottoa* (*GMROI*), joka huomioi sekä *voittomarginaalin* että *varaston kiertonopeuden* (*sales-to-stock ratio*). Mitattavana suureena on tällöin nimenomaan sijoituksesta saatava voitto aikayksikössä. GMROI määritellään seuraavasti:

$$\begin{aligned} \text{GMROI} &= \text{bruttovoittoprosentti} \times \text{varaston kiertonopeus} \\ &= \frac{\text{liikevoitto}}{\text{varaston arvo}} \end{aligned} \tag{2.1}$$

ja varaston kiertonopeus puolestaan seuraavasti:

$$\text{varaston kiertonopeus} = \frac{\text{liikevaihto}}{\text{varaston arvo}} . \tag{2.2}$$

Varaston arvo on tuotteiden hankintahinnan perusteella laskettu tarkastelujakson keskiarvo.

2.1.2 Valikoimansuunnittelu

Kun taloudelliset vaatimukset ja kriteerit on määritetty, voidaan valikoiman suunnittelu aloittaa yksittäisen tuotteiden myynnin ennustamisella. Tuotteiden myyntikäytös vaihtelee suuresti tuotteesta ja kategoriasta riippuen; toisilla tuotteilla myynti on tasaista läpi vuoden, kun taas osa tuotteista myy vain tiettynä sesonkina. Tuotteen elinkäyrään voidaan vaikuttaa mainostuksella ja hinnoittelulla.

Kun eri tuotteiden ja kategorioiden tuotto-odotukset ja myyntiarviot on määritelty, siirrytään kategorioiden keskinäisten suhteiden määrittämiseen ja kategorioiden sisällön (varsinaisten tuotteiden) valitsemiseen. Myynnissä olevaa kategorioiden yhdistelmää nimitetään *lajitelmaksi* tai *leveydeksi* (*variety, breadth*)

ja kategoriassa saatavilla olevia tuotteita *valikoimaksi* tai *syvyydeksi* (*assortment, depth*). Esimerkiksi tavaratalolla on suurempi lajitelma kuin kenkäliikkeellä, mutta vastaavasti kenkäliikkeen kenkävalikoima on kattavampi kuin tavaratalon.

Kirja listaa seuraavia huomioonotettavia asioita sopivan lajitelman ja valikoiman hakemisessa:

- myynnissä olevien tuotteiden tuottavuus,
- yrityksen filosofia tuotevalikoiman suhteen (paljon eri kategorioita ja vähän valinnanvaraa vai vähän kategorioita ja paljon valinnanvaraa),
- myymälöiden koon asettamat rajoitukset ja
- täydentävien tuotteiden myynti (henkselit ja vyöt housunostajalle).

Lisäksi on otettava huomioon tuotteiden saatavuus. Muiden tekijöiden pysyessä samana, varastoitavien tuotteiden valikoiman kasvaessa kappalemäärä pienenee, ja sitä todennäköisempää on varaston myyminen loppuun ja potentiaalisten myyntien menettäminen. Tätä mitataan usein *palveluasteella*, joka on tyydytetyn kysynnän prosentuaalinen osuus kaikesta kysynnästä.

2.1.3 Ennustaminen ja ennusteiden analysointi

Kuten edellisissä kappaleissa on tullut ilmi, myynnin ennustaminen on olennainen osa onnistunutta valikoimanhallintaa. Eri tuotteille kohdistettavat markkinointiponnistukset ovat tiukasti sidoksissa tuotteen menekkiin.

Ennustuksen laatimisen ja toteutumisen välillä on eri syistä johtuvaa *viivettä* (*lead time*), joka koostuu muun muassa seuraavista tekijöistä:

- tietojen keräämiseen ja jalostamiseen kuluva aika,
- ennustusprosessiin kuluva aika ja
- tilauksen toimitusaika.

Näistä tekijöistä johtuen valikoimanhallinnassa on loppuunmyyntitilanteiden ehkäisemiseksi varauduttava *puutetilanteisiin*, ja varattava tietty minimimäärä tuotteita *varmuusvarastoon* (*backup stock, buffer stock*).

Ylläolevat seikat huomioiden voidaan laatia valikoimasuunnitelma, joka kuvaa kunkin tuotekategorian koon ja sisällön. Suunnitelman yksityiskohtaisuus riippuu toimialasta ja sen alttiudesta muotivaihteluille. Kuvattu prosessi ja päätösten teko sen pohjalta on viime kädessä suunnitelman laatijan kokemuksesta ja ennusteiden luotettavuudesta kiinni.

Tuotehierarkian kaikilla tasoilla tehtävien muutosten apuna on eri analyysiteknikoita.

ABC-analyysi perustuu siihen, että pieni osa tuotteista tuo pääosan tuloista ("80–20" -sääntö, tai "Pareton laki"; [2] on yleisluontoinen katsaus ilmiöstä ja sen suhteesta potenssilakijakaumiin). Analyysissä tuotteet järjestetään sopivan kriteerin mukaan, kuten kokonaistuoton, kappalemääräisen myynnin tai myynnin myymäläneliömetriä kohden. Tämän jälkeen tuotteet jaetaan muutamaan ryhmään ja myyntitavoitteet ja menetelmät asetetaan kullekin ryhmälle erikseen: hyvin myyviä A-tuotteita pidetään esimerkiksi aina varastossa joka kaupassa, kun taas heikosti myyviä C-tuotteita saa vain suurimmista myymälöistä. Huonoimmin myyvät D-tuotteet voidaan jopa kokonaan jättää pois.

Myyntianalyysissä seurataan aikaisemmin tehtyjen myyntiennusteiden toteutumista ja tehdään tämän pohjalta päätöksiä valikoiman karsimisesta, hinnanalennuksista, lisätilauksista ja vastaavista toimenpiteistä. Täsmällisiä ohjeita kirja ei anna, vaan vetoaa taaskin kokemukseen ja yrityksen taloudelliseen tilanteeseen.

Monimuuttujamenetelmässä arvioidaan eri tuotteiden tai toimittajien soveltuvuutta eri mittareilla. Mahdollisina mittareina voidaan käyttää esimerkiksi taloudellisia tekijöitä, kuten tuotteesta saatavaa kateprosenttia, logistisia tekijöitä, kuten toimitusvarmuutta, tai puhtaasti tunneperustaisia syitä, kuten alkuperämaata tai muodikkuutta. Kukin mittari arvioidaan samalla asteikolla (esimerkiksi 1–10), minkä jälkeen mittareiden arvoista lasketaan painotettu keskiarvo. Käytettävien mittareiden ja painoarvojen valinta on toimiala- ja asiakaskohtaista.

2.2 Holmström (1998)

Jan Holmström kirjoittaa tutkimuksessaan [20] kysynnän ennustamisesta; tutkimuksen soveltamisalana on vähittäistavarakauppa yleisesti.

Holmström kritisoi perinteistä mallia, jossa kysyntä jaetaan peruskysyntään, trendiin ja kausivaihteluun. Tällaisissa malleissa tarkastelun kohteena on *DFU, demand forecasting unit*, joka muodostuu tarkasteltavasta yksiköstä (tuote, tuoteryhmä jne), tarkastelutasosta (kokonaiskysyntä, maantieteellinen jaottelu, ketjut, ...) ja tarkastelujaksosta.

Perinteisen tarkastelun yhtenä avainongelmana on sopivan mittakaavan löytäminen, sillä mitä alemmalle tasolle mallissa edetään, sitä epäluotettavampaa ja epätarkempaa data tyypillisesti on. Samaten liian lyhyiden tarkastelujaksojen käyttö voi johtaa tilausmäärien putoamiseen minimitoimitusrajan alle, jolloin ennustuksesta tulee binäärinen "tilaanko vai enkö tilaa" -malli. Lisäksi mallin hienojakoisuuden lisääntyessä sen tulkitsijalla on vastuullaan enemmän ja enemmän työtä.

Olettaen tuotteiden kysynnän noudattavan Pareton lakia [2] Holmström ehdottaa mallia, jossa asiantuntija tai ryhmä asiantuntijoita ennustavat 10 parhaiten myyvää tuotetta järjestettynä[1]; loput tuotteet järjestetään näiden jälkeen käyttäen hyväksi menneiden tarkastelujaksojen dataa ja mahdollista käytettävissä olevaa lisäinformaatiota. Tämän lisäksi asiantuntijalähteistä hankitaan arvio kokonaismyynnistä (tämän ennustaminen on tyypillisesti helpompaa ja luotettavampaa kuin yksittäisten tuotteiden ennustaminen).

Yksittäisten tuotteiden osuudet ("Fraction") myyntimääristä voidaan laskea Pareton laki -oletuksen nojalla tuotteiden järjestysluvusta ("rank") seuraavasti:

$$\text{Fraction(rank)} = 1/(\text{rank} + \text{constant})^{1+\text{power}} \tag{2.3}$$

Vakio "constant" ja potenssi "power" toimivat säätötekijöinä; vakion pienentäminen lisää pienen järjestysluvun tuotteiden suhteellista osuutta samaten kuin potenssin kasvattaminen.

Osuudet myyntimääristä muutetaan suoraviivaisesti arvoennusteeksi ("Value Forecast"):

$$\text{Value\_Forecast(rank)} = \text{Total\_Sales} \times \text{Fraction(rank)}/\text{Total\_Fraction} \tag{2.4}$$

"Total_Sales" on aiemmin mainittu kokonaismyynti ja "Total_Fraction" kaikkien osuuksien summa.

Tulosta kokeiltiin kentällä yhden myyntijakson (vuosineljänneksen) ajan. Huomattiin, että yksittäisiin avainasiakkaisiin keskittyvillä myyntiedustajilla oli tapana yliarvioida "omien" tuotteidensa menekki. Sen sijaan kaikesta myynnistä vastaavan myyntijohtajan järjestysennusteet olivat erittäin tarkkoja. Holmström esittää, että pidemmässä pilotissa myyntiedustajilla olisi mahdollista oppia virheistään ja tarjota tarkempia ennusteita jatkossa.

Top-10:n ulkopuolelle jäävien tuotteiden ennustamisen helpottamiseksi tutkimus huomauttaa, että vuosineljänneksen paras viikkokohtainen sijoitus korre-

[1] 10 parasta tuotetta on oletettavasti helppo tunnistaa, samaten tärkeimmät markkinointipanostukset ovat luultavasti keskitettyinä näihin tuotteisiin.

loi vahvasti (korrelaatiokerroin 0,81) koko tarkastelujakson sijoituksen kanssa. Top-10 -artikkeleiden keskinäiset sijoitukset sen sijaan vaihtelivat tutkimuksen aikana jatkuvasti joko valikoimamuutosten, hinnoittelun tai mainostuksen vaikutuksesta. Tältä pohjalta näiden tuotteiden erikoiskäsittely on perusteltua.

Kokeiden lopputuloksena perinteisen mallin ennusteet olivat hieman tarkempia kuin ehdotetun mallin. Holmströmin mukaan tämä oli odotettavissa, sillä perinteisellä mallilla oli takanaan enemmän kokemuspohjaa; myös käytetyissä datankeruuprosesseissa olisi ollut parannettavaa.

2.3 Kaija Pöysti: Diplomityö (1985)

Kaija Pöystin diplomityössä [26] tutkittiin Rautakirjan[2] toimeksiannosta aikakauslehtien irtonumeromyynnin kokonaiskysyntää. Diplomityön tavoitteena oli tarjota Lehtipisteelle ja lehtikustantajille apuvälineitä sekä normaalien lehtien tapauksissa että lehtikampanjoita suunnitellessa.

Diplomityön mukaan kustantajan ja Rautakirjan intressit poikkeavat hieman toisistaan; kustantajalle tilaajalukija on luotettavampi tulonlähde kuin irtonumero-ostaja, Rautakirja puolestaan elää myyntiprovisioista. Samaten kustantajalla on mahdollisesti käytettävissä omista lehdistään enemmän ja parempaa tietoa, kun taas Rautakirjan on pärjättävä usean eri kustantajan eri tyyppisten lehtien kanssa, mikä rajaa toimintavapautta. Pöystin diplomityö lähestyy ongelmaa nimenomaan Rautakirjan näkökulmasta.

Diplomityössä keskitytään kolmeen eri skenaarioon:

1. Kustantaja on päättänyt kokonaistoimitusmäärän, jolloin Rautakirjalla on mahdollisuus jättää jakelematta osa lehdistä säästääkseen toimituskuluissa.

2. Myyntipisteet (ja mahdolliset kampanjapisteet) on valittu ennalta, mutta toimitusmäärä on avoin.

3. Tehdään pitkän tähtäimen suunnittelua vuositasolla tulevia markkinointitoimenpiteitä varten.

Pöysti käsittelee eri tyyppisiä ennustemenetelmiä: kvalitatiivisia, autoprojektiivisia ja kausaalisia. *Kvalitatiiviset menetelmät* (*judgemental forecasts*) perustuvat subjektiivisiin määritelmiin ja asiantuntija-analyyseihin, eikä niitä

[2]Lehtipiste on osa Rautakirja-yhtymää.

näin ollen harkita tutkitun ongelman puitteissa. *Autoprojektiiviset menetelmät* (*univariate methods*) perustuvat ennustettavan muuttujan mallintamiseen muuttujan edellisten arvojen perusteella. *Kausaaliset menetelmät* (*multivariate methods*) perustuvat ennustettavan muuttujan mallintamiseen muilla muuttujilla. Hyvä perusteos autoprojektiivisista ja kausaalisista menetelmistä on Chris Chatfieldin *Time-series Forecasting* [10].

Perustuen lehtien kysynnän ominaispiirteisiin, Rautakirjan asiantuntijakommentteihin ja aikaisempiin huonoihin kokemuksiin autoprojektiivisesta aikasarja-analyysistä, diplomityö päätyy kausaalisten regressiomallien käyttöön.

Kausivaihtelun kompensoimiseksi Pöystin diplomityössä päädyttiin asiantuntijoiden laatimiin lehtikohtaisten kuukausi-indeksien käyttöön. Puutetilanteiden kompensointi tehtiin olettamalla puutteet Poisson-jakautuneeksi. Tätä oletusta tuki Pöystin siteeraama Rautakirjan sisäinen tutkimus asiasta. Jakauman parametriksi θ arvioitiin neljäsosa keskimääräisestä myynnistä, mutta Pöysti kuitenkin suosittaa kysely- tai pistokokeita θ:n sopivan arvon määrittämiseksi.

Pöysti valitsi regressiomalliinsa alustavasti kaksikymmentä selittäjää Rautakirjan toimittamasta datasta ja vertaili näiden välisiä korrelaatioita; korrelaatioiden luotettavuutta puolestaan tutkittiin t-testin avulla. Kaikille lehdille toimivaa perusselittäjäjoukkoa ei kuitenkaan löytynyt, vaan sopiva selittäjäjoukko riippui lehden luonteesta.

Lopulliseksi mallinnusratkaisuksi Pöysti valitsi lineaarisen regressiomallin kahdeksalla selittäjällä, joista kaksi oli jonkin toisen selittäjän neliö. Näin ollen varsinaisia syötemuuttujia oli kuusi. Kaikkia näitä ei kuitenkaan käytetty jokaisen lehden kohdalla, vaan toteutettu järjestelmä kokeili kaikki mahdolliset kombinaatiot läpi ja valitsi kullekin lehdelle parhaan selitysasteen antavan mallin.

Mallin toimivuudesta Pöystillä on tarjota vähän informaatiota, johtuen kelvollisen vertailudatan puutteesta. Historiadataan verrattuna ennustetarkkuus oli 80 % – 90 % välillä. Alustavan tutkimuksen mukaan hyvin pienet kampanjat ja hyvin suuret kampanjat tuottivat mallille eniten vaikeuksia. Pöystin mukaan tulos johtuu pienimmän neliösumman mukaan operoivasta regressiomallista, jolla vaihtelevassa aineistossa "ääripisteet saadaan sovitetuksi kaikkein heikoimmin."

2.4 Karlos Artto: Väitöskirja (1994)

Karlos Artto tutki väitöskirjassaan [1] suomalaisten aikakauslehtien kysynnän arvioimista myyntipistetasolla sekä myyntipistekohtaisen kappalemäärän optimaalista jakamista. Kyseisellä tutkimuksella on paljon yhtymäkohtia tässä

diplomityössä pohdittuun ongelmaan.

2.4.1 Tutkimusongelman määrittely

Arton mukaan lehtimarkkinat ovat erinomainen esimerkki häviävistä (*perishable*) tuotteista; muita samanlaisia ovat esimerkiksi nopeasti pilaantuvat ruokatuotteet tai lentoliput. Näille tuotteille on ominaista suhteellisen lyhyt myyntiaika ja varaston kokonainen tai osittainen siirtymättömyys tarkastelujaksolta toiselle (lehtien tapauksessa eri numeroiden välillä). Myös tuotteen varastointi ennen myyntiintuloa on tyypillisesti vaikeaa, ellei mahdotonta. Aihetta on tutkittu pääasiassa sovelluskohtaisesti, ilman suurempaa yleiskatsausta asiaan.

Koska Lehtipiste toimii lehtikohtaisen provision pohjalta, myyntipisteisiin toimitettavien lehtien suhteen taloudellinen riski on kustantajalla. Myymättä jääneistä lehdistä saa tyypillisesti korkeintaan kierrätyspaperikorvauksen takaisin. Tutkimuksessa oletettiin lisäksi, että kaikki tarkasteltavan lehden kappaleet toimitetaan kerralla eikä välivarastoja käytetä; reaalimaailman käytännöt tukevat tätä olettamusta.

Artto pohtii kysynnälle sopivaa todennäköisyysjakaumaa ja analysoimalla ostotapahtuman luonnetta harkitsee muun muassa binomijakaumaa ja Poisson-jakaumaa. Lisäksi Artto kokeilee myös eri jakaumien sopivuutta dataan ja päätyy lopulta olettamaan jakauman normaalijakautuneeksi. Tätä päätöstä tukevat sekä datan empiirinen analyysi että laskennallinen helppous. Lisäksi Artto viittaa aikaisempaan tutkimukseen, jonka mukaan normaalijakauma on hyvä approksimaatio, vaikka kysyntä poikkeaisikin siitä lievästi.

2.4.2 Toimitusmäärän optimointi

Varastonhallinta on eräs operaatiotutkimuksen tutkimuskohteista; häviävien tuotteiden tapauksessa ongelma esitetään usein juuri lehtien irtonumeromyynnin avulla. Kyseistä ongelmanasettelua kutsutaan *newsboy-ongelmaksi*, ja siihen on johdettu useita ratkaisuja riippuen alkuoletuksista [14].

Lehtipisteen tapauksessa ongelman formuloinnissa ei tarvitse ottaa huomioon alkuinventaariota; Artto jättää myös mahdolliset kiinteät kustannukset huomiotta, joten esitettyä mallia voi pitää newsboy-ongelman yksinkertaisimpana muotona.

Seuraavassa optimoitava suure on myyntiin otettavien lehtien kappalemäärä y. Kysynnän oletetaan noudattavan todennäköisyysjakaumaa $f(x)$, lehden keskimääräinen valmistuskustannus on c, myytyjen lehtien tuotto p ja myymättömistä lehdistä saatava palautuskorvaus u.

Maksimoimalla tuoton odotusarvo

$$E[P(y)] = p \times \left(\int_{-\infty}^{y} x f(x)\, dx + \int_{y}^{\infty} y f(x)\, dx \right) - cy$$
$$-u \times \int_{-\infty}^{y} (y - x) f(x) dx \tag{2.5}$$

saadaan ratkaisuksi

$$y_{\text{opt}} = F^{-1}(1 - (c + u)/(p + u)), \tag{2.6}$$

jossa $F(x)$ on $f(x)$:n kertymäfunktio.

Olettamalla $x \sim N(\mu, \sigma^2)$ ja sivuuttamalla palautuskustannukset u merkityksettömän pieninä Artto saa optimiratkaisuksi

$$y_{\text{opt}} = \mu + \tau\sigma, \qquad F(\tau) = F(1 - c/p)\,. \tag{2.7}$$

Näin ollen ongelmaksi jää kysynnän jakauman ennustaminen.

2.4.3 Kysynnän ennustaminen

Kysyntäjakauman ennustamisessa Artto lähtee liikkeelle tutkimalla myyntipisteeseen lähetettyjen ja sieltä palautettujen lehtien lukumääriä (y ja r). Todellisen kysynnän selvittämisessä on kuitenkin ongelmana lehtien loppuminen myyntipisteestä. Tällöin kysyntä tuli joko tyydytettyä täsmälleen, tai kysyntä x oli suurempi kuin y. Viimeeksimainitussa tilanteessa on myös saattanut käydä niin, että myyntipisteen ylijäänyt kysyntä on tyydytetty jossain toisessa myyntipisteessä, ja kokonaiskysyntä näin ollen vastaakin havaittua.

Mikäli loppuunmyyntitilanteita ei kompensoida, myyntikappaleista laskettu kysyntä vääristyy liian alhaiseksi. Tämän kompensointiin Artto harkitsee sekä Tobit-malleja, ilmoitetun myyntimäärän kertomista ennaltavalitulla vakiolla tai arvioidun kysynnän jakaumasta laskettua ehdollista odotusarvoa. Artto päätyy viimeiseen vaihtoehtoon, koska se on matemaattisesti hyvin perusteltavissa ja helposti laskettavissa tai arvioitavissa.

Artto käyttää aluksi mallinsa pohjana eksponentiaalisesti tasoitettua aikasarjaa jakauman keskiarvon (kaava 2.8) ja keskipoikkeaman (*Mean Absolute Deviation*, kaava 2.9) ennustamiseen. Seuraavassa x_t on mitattu kysyntä ja α ja β tasoitusvakioita, joille pätee $0 < \alpha < 1$ ja $0 < \beta < 1$.

$$\hat{\mu}_{t+1} = \alpha x_t + (1 - \alpha)\hat{\mu}_t \tag{2.8}$$

$$\widehat{\mathrm{MAD}}_{t+1} = \beta\,|x_t - \hat{\mu}_t| + (1 - \beta)\widehat{\mathrm{MAD}}_t \tag{2.9}$$

Keskipoikkeamasta päästään normaalijakauman tapauksessa yksinkertaisella muunnoksella keskihajontaan:

$$\hat{\theta} = \sqrt{\pi/2}\,\widehat{\mathrm{MAD}}\,. \tag{2.10}$$

Mallia tarkennettiin tämän jälkeen ottamalla huomioon lehden hinta (oliko lehti myynnissä normaalilla vai alennetulla hinnalla) sekä kausivaihtelu, jota kompensoitiin menneistä tapahtumista lasketulla kausivaihteluindeksillä. Lisäksi lyhyesti käsiteltiin, miten ennusteviive vaikuttaa ennusteisiin[3]. Tiivistettynä: mitä pidempi ennustusviive, sitä suurempi on ennustusvirheen varianssi. Tämä kuitenkin riippuu osaksi myös käytettävästä mallista.

Artto johtaa lopuksi algoritmin lehtien optimaalisen jaon muodostamiseen. Ensiksi kaavan 2.6 perusteella määritetään optimaalinen palvelutaso, josta saadaan arvioidun kysynnän jakauman kautta määritettyä haluttu jakomäärä. Mikäli kaikkien myyntipisteiden yhteenlaskettu jakomäärä vastaa toimitettavana olevien lehtien määrää, muodostettu jako on optimaalinen.

Mikäli toimitettavia lehtiä on kuitenkin enemmän kuin optimaalisuus edellyttäisi, ylimääräiset lehdet lisätään yksi kerrallaan niihin myyntipisteisiin, joista odotettavissa oleva rajatuotto on suurin. Vastaavasti, mikäli toimitettavia lehtiä on liian vähän, poistetaan jaosta lehtiä pienimmän rajatuoton mukaan.

Mallia kokeiltiin vertaamalla mallin vuosien 1986–1991 myyntihistorian perusteella vuodelle 1992 laskemia ennusteita Lehtipisteen todellisiin myyntimääriin samalta vuodelta seitsemän eri lehden osalta. Kokeissa havaittiin ensinnäkin hinnanalennusten oleva tehokas markkinointikeino: alennetulla hinnalla myytyjen lehtien kysyntä oli keskimäärin puolitoistakertainen normaalihintaisiin verrattuna. Mallin suorituskyky oli myös kiitettävä; testattujen seitsemän lehden myynti parani keskimäärin 5,4 %.

2.5 Yhteistapahtumadatan analyysi

Yhteistapahtumadatassa mielenkiinnon kohteena on kahdesta eri datajoukosta $\mathcal{X}$ ja $\mathcal{Y}$ poimittujen alkioiden x ja y samanaikainen esiintyminen. Esimerkkeinä voidaan mainita tekstidokumenttien *bag-of-words* -malli, jossa dokument-

[3]Ennusteviivettä on käsitelty tarkemmin vähittäistavarakaupan yhteydessä sivulla 6.

tia esittää sen sanojen esiintymishistogrammi. Datajoukkoina tässä ovat dokumentit ja näiden sanat. Toisena esimerkkinä voidaan mainita myös yleisesti käytetty ostoskorianalyysi, jolloin datajoukkoina toimivat asiakkaat ja ostetut tuotteet.

Yhteistapahtumamallit ovat luonteeltaan ohjaamattomia; sen sijaan tavoitteena on löytää riippuvuuksia sekä datajoukkojen väliltä että niiden sisältä. Suosittu esitysmuoto datalle on harva matriisi, jonka solussa (i, j) on alkioiden (x_i, y_j) yhteistapahtumien määrä. Datan harvuus asettaa haasteita mallinnusalgoritmeille (*zero-frequency problem* [31]). Probleema voidaan myös yleistää käsittämään useamman kuin kahden datajoukon tapaus. Tällöin ongelmaksi muodostuu niin sanottu dimensionaalisuuden kirous data-avaruuden tilavuuden kasvaessa eksponentiaalisesti ulottuvuuden mukaan, jolloin tarvitaan vastaavasti enemmän dataa luotettavien tulosten saamiseksi.

Yhteistapahtumadatasta saadaan myös mielenkiintoisia yleistyksiä ja erikoistapauksia datan esitysmuotoa vaihtelemalla. Sallimalla matriisin solujen (i, j) arvojen olevan reaalilukuja mallien sovellettavuusalue datan suhteen kasvaa, mutta osa menetelmistä voi kuitenkin menettää teoreettisen oikeutuksensa tässä tilanteessa. Vastaavasti olettamalla matriisin solut binäärisiksi (arvoina vain 0 tai 1), voidaan löytää tähän erikoistapaukseen paremmin sopivia malleja (esimerkkinä [5]).

Tässä tutkimuksessa käytettävissä olevasta aineistosta (lisätietoja kappaleessa 4) myyntitapahtumadata on malliesimerkki yhteistapahtumadatasta, lehtien ja myyntipisteiden toimiessa edellämainittuina datajoukkoina.

2.5.1 Spektraalimallit

Useat yhteistapahtumadataan sovellettavista malleista ovat saaneet innoituksensa lineaarialgebran erilaisista matriisihajotelmista. Tyyppiesimerkkeinä toimivat pääkomponenttianalyysi (PCA) [13] ja latentti semanttinen analyysi (LSA) [11]; edellinen pohjautuu ominaisarvohajotelmaan ja jälkimmäinen singulaariarvohajotelmaan. Näistä PCA on suosittu dimensionaalisuuden vähentämismenetelmä, kun taas LSA:n yhtenä sovelluskohteena on juuri tekstidokumenttien analyysi ja haku eli LSI [3].

2.5.2 PCA:n jatkokehitys diskreetin aineiston suuntaan

Lähtemällä liikkeelle PCA:n ja LSI:n probabilistisesta tulkinnasta, eri tutkijat ovat kehittäneet diskreetille aineistolle erityisesti sopivia (tyypillisesti generatiivisia) malleja. Vaihtoehtoisia malleja on useita, ja tämän tutkimuksen puit-

teissa on vaikea perehtyä niihin kaikkiin syvällisesti. Tämän osuuden viitteet ovat suunnattu asiasta kiinnostuneille lisätiedon lähteiksi.

Esimerkkinä tällaisista malleista voidaan mainita Thomas Hofmannin probabilistiset LSA-laajennukset (PLSA/PLSI) [16, 17], David Blein, Andrew Ng:n ja Michael Jordanin latentti Dirichlet-allokaatio (LDA) [6], sekä Wray Buntinen multinomiaalinen pääkomponenttianalyysi (multinomial-PCA) [8, 9]. Kuten Buntine yhteenvedossaan [8] toteaa, myös ei-negatiivisten matriisien tekijöihinjako (NMF) [24] voidaan tulkita tähän malliperheeseen kuuluvaksi siitä huolimatta, että sen alkuperäismuotoilussa [23] ei oteta kantaa probabilistiseen tulkintaan, vaan keskitytään matriisihajotelman laskemiseen ja sen tulkintaan.

On myös huomattava, että Lehtipisteellä tutkitaan tämän diplomityön jatkoprojektin puitteissa kysynnän ennustamista mPCA/LDA-sukuisella mallilla[4].

2.5.3 Komponenttimikstuurimallit

Thomas Hofmann on julkaissut yhdessä Jan Puzichan kanssa paperin [18], joka esittelee erilaisia probabilistia komponenttimikstuurimalleja, sekä johtaa kullekin näistä EM-pohjaisen [12] opetusalgoritmin. Paperin kruununjalokivi on hierarkkinen klusterimalli HACM (tunnetaan myös nimellä CAM [19, 15]), joka muodostaa hierarkkisen klusteroinnin ensimmäisen datajoukon yli, sekä klusteroi toisen datajoukon ilman hierarkiaa[5].

Valitettavasti kyseiset mallit ovat sekä muisti- että laskentatehovaatimuksiltaan raskaita [27], joten niiden käyttöä tässä projektissa ei valitettavasti voitu harkita aikataulu- ja resurssirajoitusten takia.

2.6 Yhteenveto

Kysynnän ennustamisella on suora yhteys hyvän valikoiman valitsemiseen; jos lehtikohtainen kysyntä tiedettäisiin täsmälleen, optimaalinen valikoima saataisiin napsimalla parhaat päältä. Samaan tulokseen päästäisiin, jos käytössä olisi Holmströmin ehdottama järjestysmalli. Myyntipisteessä nyt myynnissä olevien tuotteiden kysynnän ennustamiseen onkin tarjolla useita työkaluja, tässä listattujen lisäksi mm. ARIMA/Box-Jenkins -malli [7].

Tutkitussa aineistossa ei kuitenkaan ollut informaatiota diplomityön kannalta olennaiseen probleemaan: paljonko myyntipisteessä mahdollisesti aikaisem-

[4]Tämän diplomityön kirjoittaja osallistuu aktiivisesti mallin toteutustyöhön.

[5]Sekä lehdille että myyntipisteille on olemassa Lehtipisteen määrittelemä hierarkia olemassa, joten mallin soveltaminen olisi ollut helppoa.

min tuntematon lehti möisi? Vastaavasti jos jonkin (mielivaltaisen) valikoiman kokonaiskysyntä[6] voitaisiin määrittää täsmällisesti, valikoiman optimoimiseen voitaisiin käyttää erilaisia satunnaistamis- tai näytteistysalgoritmeja, kuten Gibbsin näytteistämistä tai Metropolis-algoritmia, lokaalin optimin löytämiseksi.

[6]Koko valikoiman kokonaiskysyntä voi yleisessä tapauksessa riippua lehtien välisistä suhteista; jos lehti on aiheryhmän ainoa edustaja jossain myyntipisteessä, toisen lehden lisääminen tähän aiheryhmään ei todennäköisesti kaksinkertaistaisi kysyntää.

Luku 3

Käytetyt menetelmät

Tässä luvussa esitellään lyhyesti tutkimuksessa käytettyjä matemaattisia menetelmiä.

Sananen notaatiosta: skalaarista syötettä merkitään symbolilla x. Vektorimuotoinen syöte esitetään puolestaan symbolilla $\mathbf{x} \in \mathbb{R}^n$ ja koko opetusdatan joukkoa symbolilla $\mathcal{D}$. Ennustimien antama tulos tai vastaava "ennalta tiedossa oleva" data esitetään symbolilla y tai $\mathbf{y}$. Vektorit yleisestikin esitetään lihavoituna ja matriiseja tavallisilla isoilla kirjaimilla.

3.1 Regressiomallit

Regressiossa oletetaan tulosmuuttujan y riippuvan jollain tapaa syötemuuttujista x_i, jolloin regressiomallin tyyppi määrittää riippuvuussuhteen laadun. Regressiomallia voidaan käyttää esimerkiksi selvittämään, mitkä muuttujat todellisuudessa vaikuttavat lopputulokseen tai tuntemattoman muuttujan ennustamiseen. Tyypillisesti y:n oletetaan olevan satunnaismuuttuja jollain ennalta määrätyllä jakaumalla, ja regressiomalli valitaan siten, että halutut jakauman parametrit voidaan arvioida helposti.

Kirja *Intelligent Data Analysis* [4] on hyvä johdanto regressiomalleihin sisältäen teorian lisäksi myös sovellusesimerkkejä. Seuraavat kappaleet noudattelevat kyseisen teoksen esitystä.

3.1.1 Lineaarinen regressiomalli

Lineaarinen regressiomalli olettaa tulosmuuttujan $y(t)$ arvojen olevan syötemuuttujien $x_i(t)$ ja virheen ε lineaarikombinaatio.

$$y(t) = \sum_i a_i x_i(t) + b + \varepsilon, \tag{3.1}$$

jossa a_i ovat mallin painokertoimia ja b on vakiotermi; virhe oletetaan normaalijakautuneeksi $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

On huomattava, että (yleisen) lineaarisen regressiomallin nimessä lineaarisuus merkitsee kertoimien lineaarisuutta. Näin ollen syötemuuttujina voidaan käyttää alkuperäistermejä sopivan muunnoksen jälkeen, kuten syötemuuttujien korkeamman asteen termejä.

Kaavassa 3.1 voidaan käyttää laskennallisisista ja merkintäteknisistä syistä vakiotermin asemesta ylimääräistä, vakiosyötemuuttujaa $x_b(t) := 1$, jolloin b:n korvaa painokerroin a_b.

Kyseinen yhtälö voidaan jatkokäsittelyn helpottamiseksi esittää matriisimuodossa lyhyemmin $\mathbf{y} = X\mathbf{a} + \varepsilon$. Mallin parametrien sovittamiseen voidaan käyttää neliöllistä virhefunktiota

$$\mathcal{E} = \frac{1}{2}\|\mathbf{y} - X\mathbf{a} - \varepsilon\|^2 . \tag{3.2}$$

Ottamalla virhefunktion odotusarvo ja derivoimalla se saadaan pienimmän neliösumman ratkaisuksi

$$\mathbf{a} = (X^T X)^{-1} X^T \mathbf{y} . \tag{3.3}$$

3.1.2 Yleistetty lineaarinen regressiomalli

Ylläoleva lineaarinen regressiomalli (3.1) voidaan vaihtoehtoisesti formuloida myös siten, että $y(t)$ on normaalijakautunut $N(\mu, \sigma_\varepsilon^2)$. Tällöin malli esitetään muodossa:

$$E[y(t)] = \sum_i a_i x_i(t) + b . \tag{3.4}$$

Yleistetyt lineaariset mallit muuttavat näitä oletuksia siten, että $y(t)$:n jakauman ei tarvitse olla normaalijakautunut; esimerkiksi log-lineaarisissa malleissa tavoitejakauma on yleisesti eksponentiaalisesta jakaumaperheestä.

Lisäksi syötemuuttujien lineaarikombinaatio ei ole enää suoraan jakauman estimoitava parametri, vaan se muunnetaan tarkoitukseen sopivalla monotonisella ja derivoituvalla funktiolla:

$$g\left(E[y(t)]\right) = \sum a_i x_i(t) + b\,. \tag{3.5}$$

Kaavan 3.5 mukaisia malleja ovat muun muassa edellämainittu log-lineaarinen malli, jolloin $g = \log$ ja logit-malli, jolloin $g(p) = \log(p/(1-p))$. Log-lineaarinen malli soveltuu eksponentiaalisesta jakaumaperheestä lähtöisin olevan positiivisen datan kanssa käytettäväksi, ja logit-malli on puolestaan hyödyllinen $[0, 1]$-välillä olevan datan kanssa.

Mallin kertoimien arvioiminen ei käy aivan yhtä suoraviivaisesti kuin lineaarisen mallin tapauksessa virhefunktion monimutkaisuuden vuoksi. Sen sijaan voidaan kaavassa 3.5 soveltaa molempiin puoliin funktion g:n käänteisfunktiota, jolloin tulokseksi saadaan

$$E[y(t)] = g^{-1}\left(\sum a_i x_i(t) + b\right)\,. \tag{3.6}$$

Malli on nyt saman muotoinen kuin neuroverkkotutkimuksessa käytetyt perseptronimallit, joten voidaan käyttää useita tunnettuja iteratiivisia optimointimenetelmiä [13].

3.2 Itseorganisoiva kartta (SOM)

Itseorganisoiva kartta on Teuvo Kohosen [21] kehittämä ohjaamattoman oppimisen menetelmä. Siinä syöteaineiston $\mathcal{D}$ vektorit $\mathbf{x}_i$ kuvataan kaksi- tai useampiulotteiselle säännölliselle hilalle.

Kartta muodostuu joukosta mallivektoreita $\mathbf{y}_j$, joiden välille on määritelty naapurustoriippuvuudet. Nämä naapurustoriippuvuudet ovat tyypillisesti esitettävissä pieniulotteisessa avaruudessa visualisointisyistä. Esimerkkejä käytetyistä riippuvuussuhteista ovat:

- yksiulotteisessa tapauksessa jana,
- kaksiulotteisessa tapauksessa suorakulmainen tai heksagonaalinen hila ja
- kolmiulotteisessa tapauksessa suorakulmainen hila.

3.2.1 Opettaminen

Kartan vektorit sovitetaan opetusaineistoon $\mathbf{x}_i$ soveltamalla alempana kuvattua iteratiivista prosessia jokaiselle opetusaineiston vektorille. Normaalisti opetusprosessi toistetaan samalla aineistolla useita kertoja säätäen samalla opetusparametreja. Ensimmäiseksi etsitään mallivektoreista se, joka on lähinnä opetusvektoria:

$$c = \arg\min_j \|\mathbf{x}_i - \mathbf{y}_j\| \,. \tag{3.7}$$

Edellisen kohdan "voittajavektoria" ($\mathbf{y}_c$) ja sen naapurustoa siirretään kohti opetusvektoria:

$$\forall \mathbf{y}_k : \mathbf{y}_k(t+1) = \mathbf{y}_k(t) + h_{ck}(t)[\mathbf{y}_k - \mathbf{x}_i] \,. \tag{3.8}$$

Kaavassa 3.8 $h_{ck}(t)$ on naapurustofunktio, joka on määritelty naapurustohilan ylitse. Tyypillisesti naapurustofunktio on positiivinen rajatulla alueella hilasolun c ympärillä ja 0 muualla; lisäksi naapuruston koko pienenee opetuksen edetessä.

Kartan konvergoitumisen kannalta on tärkeää, että $h_{ck}(t) \to 0$, kun $t \to \infty$. Esimerkkinä voidaan mainita gaussinen tasoitusfunktio

$$h_{ck}(t) = \alpha(t) \times \exp\left(-\frac{\|r_c - r_k\|^2}{2\sigma^2(t)}\right), \tag{3.9}$$

jossa $\alpha(t)$ on monotonisesti laskeva opetuskerroin, r_c ja r_k ovat solujen koordinaatit hila-avaruudessa ja $\sigma(t)$ tasoitusfunktion (myös monotonisesti laskeva) leveysparametri.

Mallivektoreiden alkuarvot voidaan alustaa satunnaisesti, mutta kartan konvergoituminen on nopeampaa, mikäli alustaminen tapahtuu tasaisesti ja järjestelmällisesti oletetun kohdeavaruuden ylle.

On myöskin huomattava, että SOM-kartta voi olla hyvin herkkä syötedatan skaalaukselle riippuen karttaan valitusta hilarakenteesta ja käytetystä naapustofunktiosta. Kartta pyrkii asettumaan data-avaruudessa datan pääakseleiden mukaisesti, jolloin eri mittakaavassa olevat arvot dominoivat prosessia. Tyypillisesti tällaisissa tilanteissa data normalisoidaan vähentämällä muuttujista datan keskiarvo ja jakamalla tulos keskihajonnalla.

3.2.2 Käyttökohteita

Mikäli käytettävissä on opetusdatan lisäksi luokiteltua testausdataa, kartan solut voidaan merkitä opetuksen jälkeen luokkatunnuksin laskemalla kullekin testausvektorille lähin mallivektori. Kun testausdata on saatu jaettua kartan soluille, kullekin solulle määrätään näiden vektoreiden yleisin luokkatunnus. Näin kartta toimii jatkossa luokittimena.

Yllä olevaa menetelmää voidaan laajentaa siten, että luokkatunnusten sijaan testausaineistolle on määritelty skalaari- tai vektorimuotoinen ennustettava suure. Tällöin karttaa voidaan käyttää datan visualisointiin tai ennustamiseen.

Jo opetusdatan jakaminen soluille muodostaa opetusdatalle solutason klusteroinnin. Tyypillisesti kuitenkin yhden solun läheisyydessä on samankaltaisia soluja, joten vertailemalla mallivektoreiden keskinäisiä etäisyyksiä ja yhdistämällä toisiaan lähellä olevia soluja, voidaan aineisto klusteroida korkeammalla tasolla.

3.3 k:n lähimmän naapurin menetelmä (k-NN)

Jotta k:n lähimmän naapurin menetelmän toiminta selviäisi mahdollisimman hyvin, tässä työssä kuvataan sen toiminta ensin yksinkertaisen luokitteluprobleeman yhteydessä. Tämän jälkeen algoritmia laajennetaan kattamaan myös tutkimuksessa käytetty ennustamisvariantti.

Käytettävänä opetusaineistona on joukko vektoreita $\mathbf{x}_i$ ja näihin liittyviä luokkatunnuksia $\mathcal{C}_i$. Määritettäessä ennestään tuntemattoman vektorin $\mathbf{x}$ luokkaa valitaan aineistoista k lähintä vektoria $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$ ongelmaan soveltuvan etäisyysfunktion $d(\mathbf{x}, \mathbf{y})$ mukaan. Perinteisiä valintoja etäisyysfunktioksi ovat esimerkiksi vektoreiden välinen euklidinen etäisyys tai vektoreiden pistetulo.

Vektorin $\mathbf{x}$ luokka valitaan vektoreiden $\mathbf{x}_i$ luokista enemmistöperiaatteella; tarkemmin sanottuna vektorin luokaksi asetetaan se luokka $\mathcal{C}$, johon kuuluu eniten vektoreita $\mathbf{x}_j$.

Mikäli useampi luokka on "eniten edustettuna", voidaan näiden välillä valita joko ottamalla seuraavaksi lähimpiä vektoreita opetusjoukosta kunnes tasapelitilanne ratkeaa tai arpomalla voittaja luokkien esiintymistodennäköisyyksien suhteessa, mikäli tämä on tiedossa. Kahden luokan tapauksessa tilanteen voi estää valitsemalla parittoman k:n.

Mikäli data on sopivan tasaisesti jakautunut, voidaan myös k:n lähimmän vektorin sijasta valita kaikki etäisyydellä r tai sitä lähempänä olevat opetusvektorit ja niiden luokat. Näitä kahta lähestymistapaa voidaan kutsua nimillä

vakiomassa ja *vakiotilavuus.*

Oletetaan nyt, että meillä on luokkatunnusten sijaan opetusvektoreihin liittyviä muuttujia $\mathbf{y}_i$. Tällöin algoritmia voidaan käyttää ennustamiseen seuraavasti: valitaan k lähintä vektoria ja niihin liittyvät muuttujat $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$. Näiden muuttujavektorien pohjalta voidaan sitten muodostaa ennustevektori $\mathbf{y}$ yhdistämällä vektorit käsillä olevaan ongelmaan soveltuvalla menetelmällä, esimerkiksi keskiarvoistamalla:

$$\mathbf{y} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{y}_i \,. \tag{3.10}$$

Luku 4

Myyntitapahtumadata ja taustatiedot

Lehtipisteen toimittama aineisto jakautui neljään eri tietolähteeseen: lehtien taustatiedot, myyntipisteiden taustatiedot, lehtien numerotiedot sekä nämä kolme yhteennitova myyntitapahtuma-aineisto. Lisäksi myyntipisteiden taustatietoihin liitettiin postinumeron perusteella Xtract Oy:n Xtract Consumer LifeCycles -data.

Lehti- ja myyntipisteaineisto kuvaavat tilannetta sellaisena kuin se oli elokuussa 2003. Myyntitapahtumia aineistossa on vuoden 2000 alusta noin vuoden 2003 elokuuhun saakka[1].

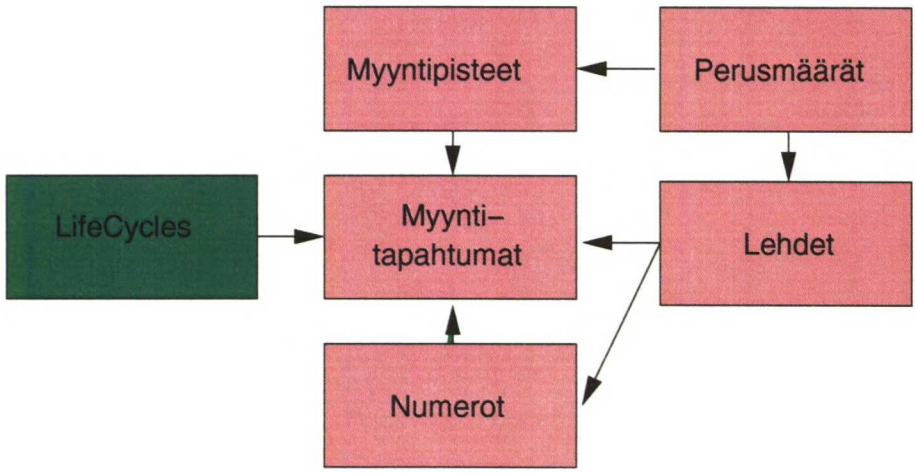
4.1 Aineiston yksityiskohdat

Tässä tutkimuksessa *lehti* on säännöllisesti ilmestyvä aikakauslehti; kustakin lehdestä ilmestyy vuodessa lehdestä riippuen vakiomäärä *numeroita*. Lehtiä myydään *myyntipisteissä*. Eri tietolähteiden suhteet on kuvattu kuvassa 4.1.

Myyntipisteisiin toimitetusta lehtien määrästä käytetään termiä *perusmäärä*. Jokaista myyntipisteessä myytyä lehden numeroa kohti muodostetaan *myyntitapahtuma*, joka kertoo, montako kappaletta lehteä myytiin ja millä hinnalla.

Kappaleen 2.1 mukaisesti myös Lehtipisteen tuotteet (tässä tapauksessa lehdet) on järjestetty hierarkisesti. Hierarkian kolmen alimman tason nimet ovat *tuoteryhmä*, *aiheryhmä* ja *lehti*. Tässä tutkimuksessa keskitytään vain aikakauslehtien tuoteryhmään, joten sen ja sitä ylempien tasojen merkitys on vä-

[1] Lehtien ilmestymisrytmistä riippuen viimeisin aineistossa olevan numeron ilmestymiskuukausi vaihtelee huhtikuun ja elokuun välillä.



Kuva 4.1: Mallinnuksessa käytettävät tietolähteet ja niiden väliset riippuvuussuhteet. Punaiset tietolähteet ovat Lehtipisteen toimittamia, vihreät Xtract:in.

häinen.

Kustakin aineistokuvauksesta on jätetty pois kenttiä, mikäli kyseinen kenttä on redundantti (esimerkiksi toisen kentän tekstiselite), tai sillä ei voi katsoa olevan minkäänlaista vaikutusta jatkotoimenpiteisiin.

Muuttujien kuvauksissa käytetään seuraavia muuttujatyyppejä:

- **ID** Tunnistemuuttuja, jota käytetään eri datalähteiden yhdistelyssä.
- **C** Kategoria- eli luokkamuuttuja; eri kategorioiden määrä on tyypillisesti ilmoitettu seuraavasti: C=7.
- **B** Binäärimuuttuja (kyllä/ei-muuttuja, kaksiarvoinen kategoriamuuttuja)
- **K** Kokonaislukumuuttuja.
- **X** Liukulukumuuttuja.

4.1.1 Lehtien taustatiedot

Lehtipisteen toimittamassa aineistossa oli 6733 lehteä, joista *aktiivisia* (myynnissä ja edelleen ilmestyviä) oli 2004. Näistä rajattiin pois kertajulkaisut ja

vain kerran vuodessa ilmestyvät lehdet ja aiheryhmät "kartat/kalenterit/CD-romit" ja "keräilytuotteet"[2]. Näin ollen käytettävissä oli 1498 lehden tiedot.

Kustakin lehdestä käytettävissä olevat tiedot on kuvattu taulukossa 4.1.

4.1.2 Myyntipisteiden taustatiedot

Toimitetussa aineistossa oli 28229 myyntipistettä, joista aktiivisia ja valikoimaluokan perusteella aikakauslehtiä myyviä 5009 kappaletta. Myyntipisteistä käytettävissä oleva data on kuvattu taulukossa 4.2.

Muuttujien "kielitunnus" ja "ruotstehop" (ruotsinkielinen tehopiste) keskinäisistä suhteista on mainittava, että "ruotstehop" on varsinaisesti määritelty vain kielitunnuksen mukaan suomenkielisille myyntipisteille.

4.1.3 Numerotiedot

Numerotiedosto listaa kullekin lehdelle kaikki tarkastelujaksolla ilmestyneet *numerot* (ID), näiden *hinnan* (X), *ilmestymispäivämäärän* ja *palautusviikon* (K). Näistä hintakenttään pätevät samat varaukset kuin aiemmin lehtitietojen yhteydessä mainittiin. Ilmestymispäivämääränä myyntipiste saa laittaa lehden esille myytäväksi, ja lehdet kerätään myynnistä viimeistään palautusviikkoa edeltävän viikon lopussa. Näitä tietueita on aineistossa 52703 kpl. Osa näistä tietueista saattaa liittyä aiemmin poisrajattuun aineistoon, mutta kyseiset tietueet ovat tällöin karsiutuneet tietojen yhdistelyn aikana.

4.1.4 Myyntitapahtumat

Aineistossa on noin 34 miljoonaa myyntitapahtumaa. Yksi myyntitapahtuma kertoo yhden lehden yhden numeron kappalemääräisen myynnin yhdessä myyntipisteessä. Lisäksi aineistossa on kyseisessä myyntipisteessä kyseiselle numerolle käytetty myyntihinta, joka on tarkin käytettävissä oleva tieto lehden hinnoittelusta.

[2]Aineistossa oli mukana myös joitain aikakauslehtien tapaan myytäviä tuotteita, kuten sarjakuva-albumeita, jotka eivät tiukasti asiaa katsoen ole lehtiä. Ne kuitenkin sisällytettiin tutkimukseen.

Muuttuja	**Kuvaus**	**Tyyppi**
Lehtikoodi	Lehden tunnistekoodi	ID
Lehtinimi	Lehden nimi	ID
Lehtihinta	Lehden perusmyyntihinta. Huomaa, että saman lehden saman numeron hinta voi vaihdella eri myyntipisteiden välillä kampanjaperusteisesti.	X
Lehtiryhmä	Tutkimuksessa vakio ("aikakauslehdet")	C
Tuoteryhmä	Lehtien korkean tason jaottelu	C=10
Aiheryhmä	Matalamman tason jaottelu (tuoteryhmän tarkempi jaottelu)	C=32*
Alkuperämaa	Lehden julkaisumaa	C†
Numtun	Ilmestyvätkö lehden numerot numerojärjestyksessä vai eivät?	B
Toimno	Lehden kustantajan tunnus	ID
Kieli	Lehden kieli	C‡
Tilatunnus	Lehden tila; aktiivinen vai ei?	B
Ilmkerratvv	Ilmestymiskerrat vuodessa. Kertajulkaisuilla ja kerran vuodessa ilmestyvillä julkaisuilla molemmilla kentän arvona 1.	K
Tilausviive	Kuinka paljon ennen toimitusta lehti täytyy tilata kustantajalta (päivissä)	K
Kassalehti	Myydäänkö lehteä kassojen lähellä olevissa telineissä?	B

*Katso myös taulukko B.3.
†Katso myös taulukko B.1.
‡Katso myös taulukko B.2.

Taulukko 4.1: Lehtien muuttujat

Muuttuja	Kuvaus	Tyyppi
Mypikoodi	Myyntipisteen tunnistekoodi	ID
Mypinimi	Myyntipisteen nimi	ID
Ketju	Mihin ketjuun myyntipiste kuuluu?	C $\approx$ 50
Toimiala	Mikä on myyntipisteen toimiala?	C $\approx$ 20*
Keskusliike	Minkä keskusliikkeen (K-ryhmä, S-ryhmä) alaisuudessa myyntipiste on?	C
Postinumero	Myyntipisteen postinumero	ID
Kuntano	Myyntipisteen kunnan (KELA-)numero	ID
Pokkaripiste	Minkätasoinen taskukirjavalikoima myyntipisteessä on	C=5
Valikoimaluokka	Mitä tuotteita myyntipiste myy?	C
Mypitilatunnus	Onko myyntipiste aktiivinen?	B
Kielitunnus	Onko myyntipiste ruotsinkielinen?	B
Ruotstehop	Myykö myyntipiste paljon ruotsinkielisiä lehtiä?	B
Pvtavmyynti	Päivittäistavaramyynti miljoonina euroina	X
Sesonkipiste	Muuttuuko myyntipisteen myynti kesällä ratkaisevasti	B
Lähiosoite	Osoite	ID
Pintala	Pinta-ala neliömetreinä	X
Kassalkm	Kassojen lukumäärä	K
Myyntipiiri	Lehtipisteen käyttämä maantieteellinen jaottelu	C=22

*Katso myös taulukko B.4.

Taulukko 4.2: Myyntipisteiden muuttujat

4.1.5 Nykyinen valikoima

Lehtipiste toimitti lisäksi ennusteiden laskemista varten tiedot lehtien perusmääristä. Tästä tiedosta voitiin helposti selvittää myyntipisteiden nykyiset valikoimat, sillä perusmäärätiedostossa oli näille myyntipiste—lehti -pareilla nollasta poikkeava arvo.

Äskettäin myyntipisteen valikoimista heikon kysynnän vuoksi poistetut lehdet olivat myös tässä tiedostossa mukana, mutta näiden perusmäärä oli merkitty nollaksi.

Perusmäärätietoja käytettiin myyntipisteissä tapahtuvien muutosten seurantaan, ja samalla estettiin vastikään poistettujen lehtien esiintyminen ennusteissa.

4.1.6 Xtract Consumer LifeCycles

Xtract Consumer LifeCycles -aineistossa [32] on laskettu 250 m × 250 m kokoisille ruuduille useita demografisia mittareita. Tämä aineisto on myös aggregoitu yhteen postinumerotasolle, ja tämän tason dataa käytettiin mallinnuksessa. Koska useilla muuttujilla oli keskinäisiä riippuvuuksia, aineiston ulottuvuutta vähennettiin pääkomponenttianalyysillä ja syntyneistä muuttujista käytettiin mallinnuksessa kymmentä suurimmalla ominaisarvolla varustettua vektoria.

4.1.7 Rajaukset

Yhteenvetona aineistosta otettiin mukaan seuraavat myyntipisteet ja lehdet sekä näihin liittyvät numero- ja myyntitapahtumatiedot:

- aikakauslehtiä myyvät aktiiviset myyntipisteet ja
- useammin kuin kerran vuodessa ilmestyvät aktiiviset aikakauslehdet.

Lisäksi ennusteita laadittaessa tietyt aiheryhmät ja tuotteet sivuutettiin vaihtelevista syistä; tarkempi kuvaus on kappaleessa 5.4 (s. 41).

4.2 Normalisointi

Kuten edellä on selitetty, myyntitapahtumat on esitetty aineistossa numerokohtaisesti. Koska lehden myyntiaika vaihtelee lehdestä riippuen viikosta vuoteen, eri lehtien numerokohtainen vertailu ei ole kannattavaa. Lisäksi lehden

ostotodennäköisyys riippuu lehden uutuudesta; ihmiset ostavat mieluummin vastailmestyneen lehden kuin esimerkiksi noin kuukauden vanhan lehden. Lisäksi uusia lehtiä ja myyntipisteitä tulee jatkuvasti mukaan järjestelmään samalla, kun vanhoja poistuu.

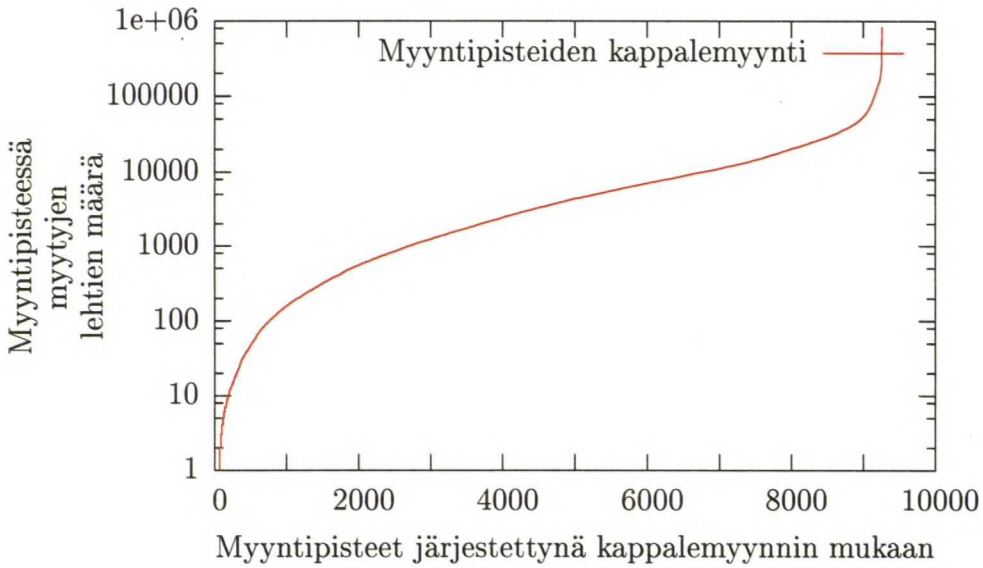
Kyseinen ongelma ratkaistiin tutkimalla kunkin lehden keskimääräistä euromääräistä viikkomyyntiä sopivalla tarkastelujaksolla, jolloin eri lehtien välinen tai lehden sisäinen vaihtelu ei vaikuta tuloksiin. Lisäksi kyseinen mittaussuure on liiketaloudellisesti hyvin perusteltavissa oleva optimointikriteeri.

Toisaalta tutkittava ongelma on luonteeltaan jokseenkin epästationäärinen, jolloin liian pitkä keskiarvoistuskausi voi kätkeä tärkeää informaatiota markkinoiden muutoksesta. Teimme tämän vuoksi lisäoletuksen, että ongelma on vuoden aikaskaalalla likimain stationäärinen ja tällainen keskiarvoistus voidaan tehdä.

Toinen lisähaaste normalisoinnissa olivat lehdet, joista oli useampia numeroita myynnissä samaan aikaan. Vastaavasti osaa lehdistä ei välttämättä myyty yhtäjaksoisesti, vaan numeroiden välissä oli viikkojen tauko. Näin ollen suoraviivainen myyntiajan laskeminen ei onnistunut, vaan ratkaisuksi tuli myyntiajan laskeminen jokaiselle lehdelle ja myyntipisteelle erikseen seuraavasti:

- Numeroille on annettu numerotiedostossa ilmestymispäivämäärä, josta saadaan ilmestymisviikko pääteltyä helposti.
- Numeroille on annettu myös palautusviikko.
- Lehden myyntiviikot ovat kaikki viikot ilmestymisviikon ja palautusviikon välillä ilmestymisviikko mukaan lukien, mutta palautusviikko pois lukien. Mikäli ne ovat samat, myyntiajaksi asetetaan yksi viikko.
- Osalla numeroista palautusviikkokenttää oli käytetty koodina erikoistilanteiden merkitsemiseen. Kyseisiä numeroita ei otettu mukaan tarkasteluun.

Kunkin myyntipisteen kullekin myynnissä olleelle lehdelle voitiin nyt yllämääritellyn viikkodatan ja myyntitapahtumadatan perusteella määrittää myynnissäoloviikot, ottamalla kaikkien myynnissä olleiden numeroiden myyntiviikkojen yhdiste. Koska meitä kiinnosti vain lehtien myynnissäoloaika, yhdisteen sisältö ei sinänsä ole relevantti, vain sen koko.



Kuva 4.2: Myyntipisteiden kappalemyynnin jakautuminen eri myyntipisteille

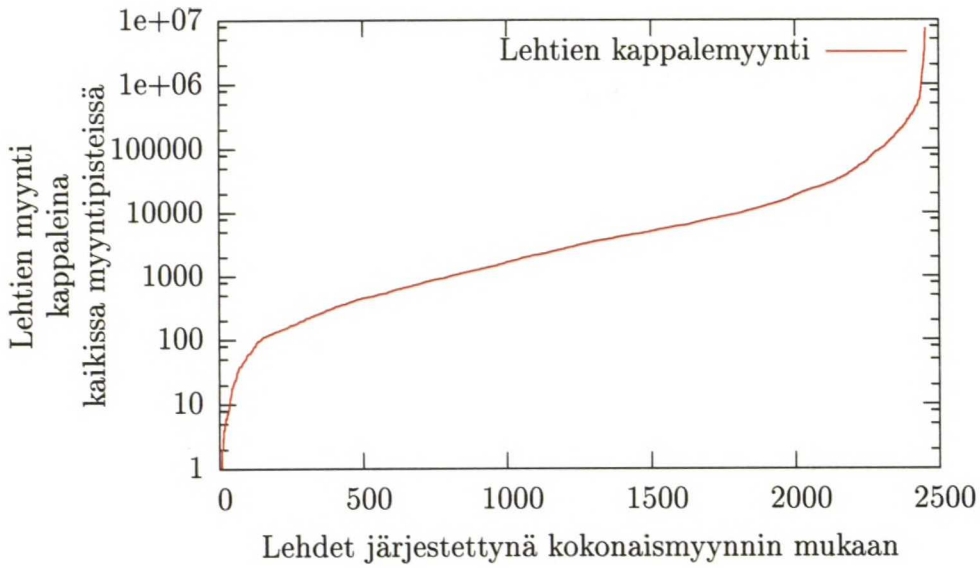
4.3 Kappalemäärien jakaumista

Kuvat 4.3 ja 4.3 esittävät myyntipisteiden ja lehtien myyntimäärien jakautumista koko 3,5 vuoden aineistosta. Myyntimäärät asettuvat käyrän keskivaiheilla likimain suoralle logaritmisella asteikolla, mikä viittaa yleensä potenssilain mukaiseen jakaumaan. Molempien käyrien ylä- ja alapäät poikkeavat kuitenkin huomattavasti suoralta; efekti johtuu osittain siitä, että lehtien ja myyntipisteiden valikoima muuttuu jatkuvasti. Osa myyntipisteistä tai lehdistä lopettaa toimintansa tarkastelujakson aikana, ja uusia tulee koko ajan lisää. Nämä vääristävät käyrän alapäätä. Yläpään vääristyminen johtuu puolestaan noin kymmenestä hyvin suuresta myyntipisteestä, joiden osalta myynti on kertaluokkaa suurempi.

4.4 Esikäsittely

Aineistolle ei juurikaan tarvinnut suorittaa esikäsittelyä, koska lähes kaikki mallinnuksessa tarvittava data oli saatavissa numeerisessa muodossa, eikä puuttuvia arvoja juurikaan ollut. Poikkeuksena mainittakoon lehtien osalta kielisyys- ja alkuperämaakentät, joissa oli numeeristen arvojen sijaan symboliset arvot. Nämä olivat käsiteltävissä yksinkertaisella 1-of-C -koodauksella.

Myyntitapahtumissa oli jonkin verran semanttisesti virheellisiä rivejä (negatii-



Kuva 4.3: Lehtien myyntimäärien jakautuminen eri lehdille

vinen myyntimäärä tai myyntimäärä puuttui), jotka loppujen lopuksi jätettiin pois ongelman rajoittuessa pääasiassa muutamaan toimintansa jo lopettaneeseen myyntipisteeseen. Näitä tietueita oli noin 21000 tai 0,062 % koko kaikista myyntitapahtumista. Lisäksi numerotiedostossa osalle lehdistä oli ilmoitettu palautusviikko, joka oli ennen myyntiintulopäivämäärää, tai erikoiskäsittelyä merkitsevä koodi "209999" tai "00000". Näitä tapauksia oli noin 900 kpl tai 1,7 % kaikista numeroista.

Luku 5

Optimaalisen valikoiman muodostaminen

Tässä diplomityössä käsitellyn projektin perusongelmana on se, että käytettävissä ei ole sopivaa opetusaineistoa. Vaikka saatavilla onkin usean vuoden myyntitapahtumatiedot, niistä on vaikea määritellä haluttua vastetta. Näin ollen ongelman tutkimiseen on paneuduttava ohjaamattoman oppimisen näkökulmasta. Mallien toimivuutta tutkitaan simulomalla historiadataa vasten siinä määrin kuin se on mahdollista, ja lopullinen toimivuus selvitetään kenttäkokeilla.

Yllä oleva huomioon ottaen projektin tavoitteeksi otettiin uusien valikoimaehdotusten laatiminen rajatulle joukolle myyntipisteitä. Tämän saavuttamiseksi pyrittiin etsimään mallia, joka joko estimoisi valikoiman kokonaismyynnin suoraan tai vaihtoehtoisesti estimoisi yksittäisten lehden myynnin, jolloin kokonaismyynti saataisiin laskemalla eri lehtien myyntiestimaatit yhteen.

Jälkimmäinen formulointi ei ota kuitenkaan huomioon lehtien vaikutusta toisiinsa; aiheryhmän osuuden kasvattaminen sen todellista kysyntää suuremmaksi tuskin luo lisämyyntiä.

5.1 Segmentointi

Projektin alkuvaiheessa pyrittiin tutustumaan dataan ja löytämään sopivat muuttujat ja menetelmät jatkokäsittelyä varten. Tästä syystä myyntipisteitä ja niihin liittyvää dataa päätettiin analysoida SOM-kartalla ja segmentoida syntynyt kartta siten, että segmenttejä voisi käyttää mahdollisesti lisäinformaationa myöhemmässä vaiheessa.

SOM-kartassa ja segmentoinnissa käytetyt syöttömuuttujat jakautuivat kahteen ryhmään; käytösperustaisiin muuttujiin ja myyntipisteen taustamuuttujiin. Käytösperustaisiin muuttujiin kuului eri aiheryhmien suhteellinen jakauma myyntipisteessä ja 15 valtakunnallisesti eniten myydyn lehden myynnin suhteellinen osuus myyntipisteessä.

Taustamuuttujiin kuului myyntipisteisiin liittyvää dataa (katso kappale 4.1.2) ja Xtract Consumer LifeCycles -aineisto. Koska Xtract Consumer LifeCycles -aineiston muuttujat korreloivat keskenään voimakkaasti, aineisto esikäsiteltiin pääkomponenttianalyysin (PCA, [13]) avulla.

SOM-kartan opettamiseen käytettiin Matlabin SOM Toolbox -ohjelmistoa [29]. Koska tarkoituksena ei ollut tutkia SOM-kartan käyttäytymistä eri opetusparametreilla, vaan saada yleiskäsitys datan luonteesta ja sen käytöksestä, kartta opetettiin SOM Toolboxin vakioasetuksilla määrittelemättä erikseen kartan topografiaa tai muita parametreja.

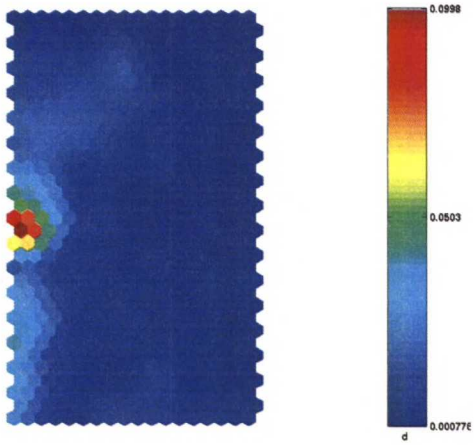
SOM-kartta opetettiin useita kertoja vaihdellen eri muuttujaryhmien välisiä painoja ja kokeillen eri normalisointitekniikoita syötemuuttujille. Syntynyttä karttaa dominoivat joko käytösperustaiset muuttujat tai taustamuuttujat riippuen muuttujien välisistä painotuksista.

Ongelman ratkaisemiseksi teimme päätöksen keskittyä mallinnuksessa pelkästään käytösperusteisiin muuttujiin, koska käytettävissä olevien taustamuuttujien määrä ja sisältö eivät SOM-kartan opettamisen yhteydessä vaikuttaneet kovinkaan informatiivisilta tietolähteiltä. Sen sijaan käytösperusteisissa muuttujissa oli havaittavissa rakennetta. Taustamuuttujia käytettiin segmenttejä kuvaavina taustatietoina keskiarvoistamalla samaan karttasoluun tulevien taustamuuttujien arvot.

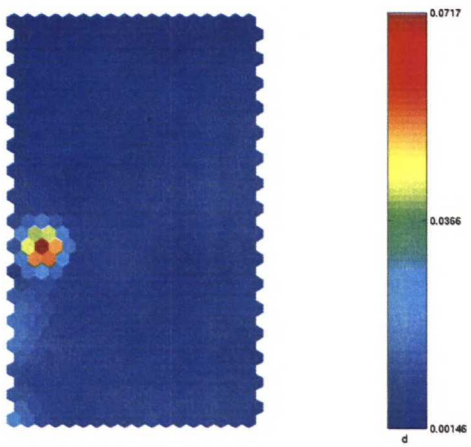
Kuvassa 5.1 esitetään kolmen aiheryhmän suhteellinen osuus kartan eri soluihin kuuluvissa myyntipisteissä. Kyseiset aiheryhmät ovat selkeästi parhaiten edustettuina rajatussa joukossa myyntipisteitä ja muualla niiden osuus on vähäisempi.

Kuvassa 5.2 esitetään puolestaan tyypillisesti miehiin yhdistettyjä aiheryhmiä. Kartalla on selkeästi huomattavissa samanmuotoinen alue, jossa myynti on muita myyntipisteitä suurempi. Etenkin auto- ja moottorilehtien (kuvat 5.2(a) ja 5.2(c)) samankaltaisuus on huomattavaa. Myös parhaiten myyvät myyntipisteet ovat joka kuvassa lähellä toisiaan.

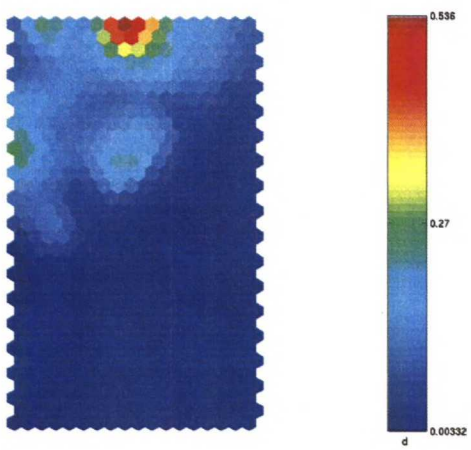
Kuvassa 5.3 nähdään puolestaan miten toimialoiltaan samanlaiset myyntipisteet sijoittuvat kartalle toistensa läheisyyteen. Kuvassa esitetään eri toimialojen suhteellista osuutta kunkin solun myyntipisteistä. On myös erityisesti huomattava, miten myyntipisteiden koon kasvaessa parhaiten edustetut alueet sijoittuvat toistensa viereen.



(a) Uutis-, talous- ja tiedelehdet

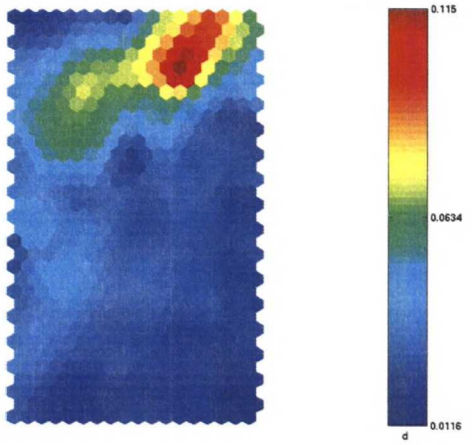


(b) Ruoka ja juhla

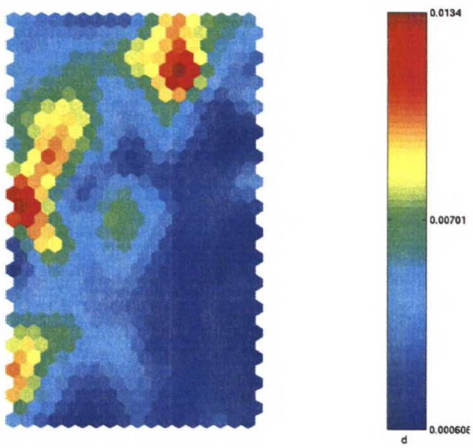


(c) Erotiikka

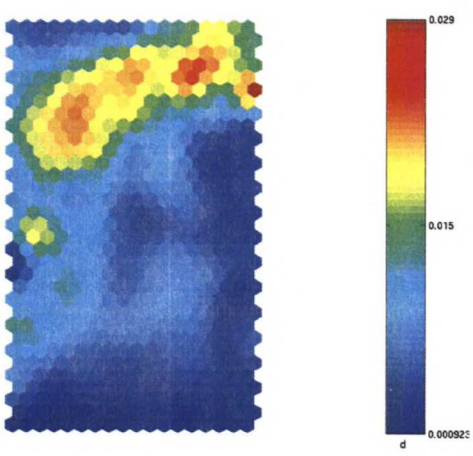
Kuva 5.1: Tiiviisti ryhmittyneet aiheryhmät. Kuvien asteikkoon voidaan liittää todennäköisyystulkinta.



(a) Autolehdet

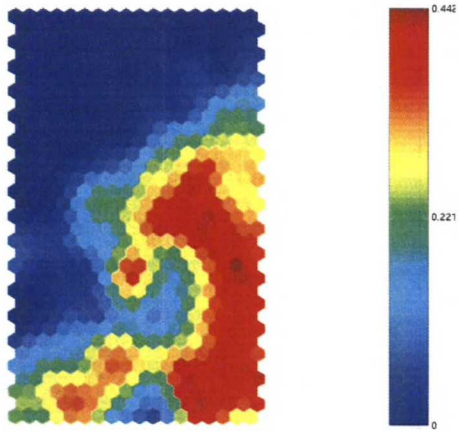


(b) Venelehdet

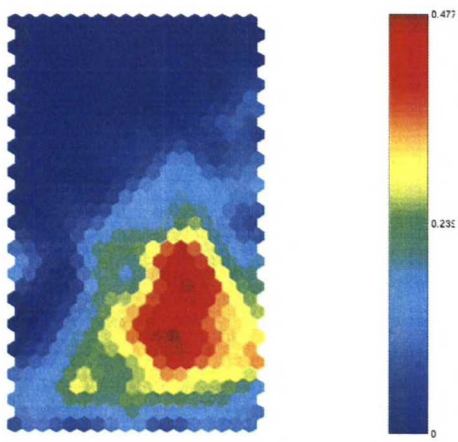


(c) Moottorilehdet

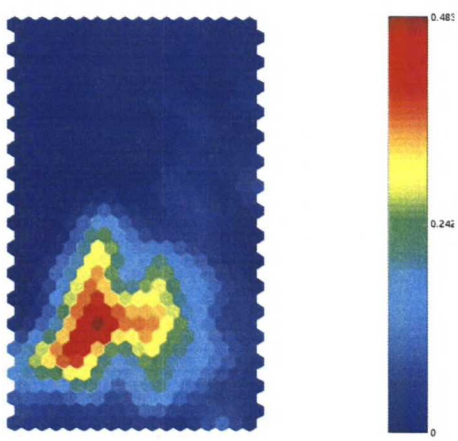
Kuva 5.2: "Miehekkäät" lehdet. Kuvien asteikkoon voidaan liittää todennäköisyystulkinta.



(a) Iso valintamyymälä

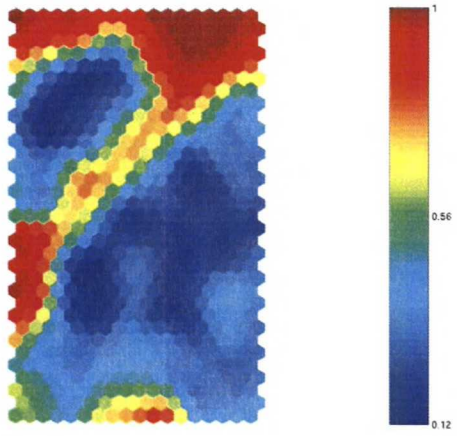


(b) Pieni supermarket

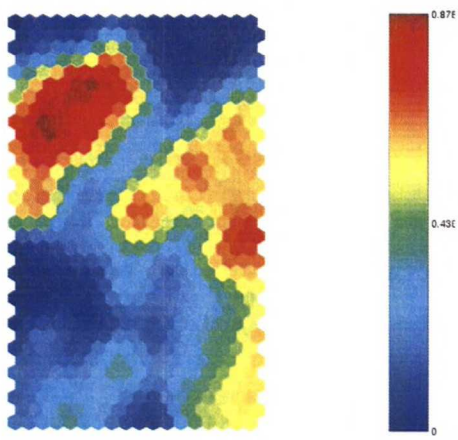


(c) Iso supermarket

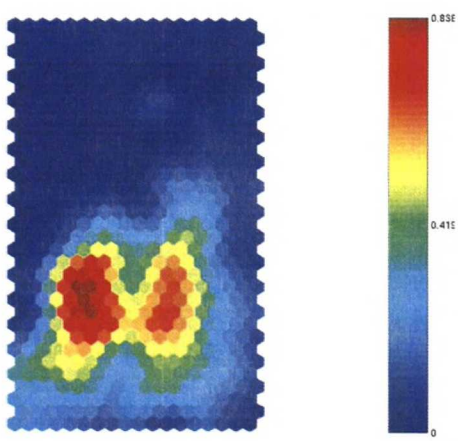
Kuva 5.3: Toimialojen jakautuminen kartalla. Kuvien asteikkoon voidaan liittää todennäköisyystulkinta.



(a) Pieni päivittäistavaramyynti



(b) Keskisuuri vähittäistavaramyynti



(c) Suuri vähittäistavaramyynti

Kuva 5.4: Päivittäistavaramyynnin jakautuminen. Kuvien asteikkoon voidaan liittää todennäköisyystulkinta.

Kuva 5.5: Muodostettu segmenttijakauma

Tätä ilmiötä kuvastaa myös kuva 5.4, jossa muuttuja "päivittäistavaramyynti" on jaettu kolmeen luokkaan, ja näiden luokkien suhteelliset osuudet esitetty kartalla.

Näiden kuvien perusteella suhteelliset aiheryhmäosuudet ovat hyvä indikaattori myyntipisteiden samankaltaisuudesta ja hyvä mallinnuksen kohde.

SOM-kartta segmentoitiin tämän jälkeen SOM Toolboxin avulla (tarkka algoritmi on kuvattu Juha Vesannon väitöskirjassa [30]), minkä jälkeen syntyneet 37 segmenttiä yhdistettiin vielä ylemmän tason segmentteihin päivittäistavaramyynnnin määrän nojalla. Segmentit on esitetty kuvassa 5.5.

5.2 Mallinnusmenetelmän valinta

Segmentoinnin muodostamisen jälkeen tutkittiin, miten ja millaisilla työkaluilla ongelmaa kannattaa ratkaista. Koska ongelma oli suhteellisen laaja ja vaikeasti hahmotettavissa kerralla, päätettiin mallien toimivuutta datan kanssa kokeilla yksinkertaisemmilla koeongelmilla.

Koska asiakas oli esittänyt erityistä mielenkiintoa myyntipisteiden kielikysymysten käsittelyyn, sopivana koeongelmana pidettiin myyntipisteen kielija-

kauman ennustamista (suomi vs. ruotsi). Ennustettavaksi muuttujaksi otettiin myyntipisteen ruotsinkielisten lehtien euromääräisen myynnin osuus koko myynnistä ja selittäviksi muuttujiksi aiheryhmien suhteellinen jakauma. Ensimmäisinä malleina olivat lineaarinen ja logit-regressiomalli opetettuna ensin koko aineistolla ja sitten kunkin segmentin sisällä erikseen.

Tämän jälkeen mallien suorituskykyä arvioitiin laskemalla regressiomallien virheen neliön keskiarvo. Tulokset ovat taulukossa 5.1. Tulokset eivät olleet mitenkään rohkaisevia, sillä ennusteiden virhe suhteessa ennustustulokseen oli parhaallakin mallilla varsin suuri. Ottaen huomioon, että ennustettavaa muuttujaa voidaan pitää ruotsinkielisten lehtien ostotodennäköisyytenä, parhaimmankin mallin virhe on $\sqrt{0{,}0291} = 17\,\%$.

Neliövirheen keskiarvo	Globaali malli	Lokaalit mallit
Lineaarinen	0,0907	0,0314
Logit-malli	0,0449	0,0291

Taulukko 5.1: Yksinkertaisten regressiomallien suorituskyky myyntipisteen kielijakauman ennustamisessa

Koska segmenttipohjaiset tulokset osoittivat selvää parannusta ja SOM-kartan kuvista oli tunnistettavissa naapurustoriippuvuuksia myös opetusaineistoon kuulumattomilla muuttujilla, päätettiin kokeilla k-NN -pohjaista ratkaisua. Kokeissa käytettiin euklidista etäisyysfunktiota sen yksinkertaisuuden ja ymmärrettävyyden takia. Teoriassa myös Kullback-Leibler -divergenssi olisi ollut mahdollinen, koska muuttujina käytetyt suhteelliset aiheryhmäjakaumat voidaan periaatteessa tulkita lehtien todennäköisyysjakaumina, joiden eroavaisuuksien vertailuun KL-divergenssi on luotu. KL-divergenssi on kuitenkin epäsymmetrinen, eikä sitä katsottu tarpeelliseksi kokeilla käytössä olleen aikataulun puitteissa.

Tässä tapauksessa naapurustoon otettiin 20 lähintä myyntipistettä, ja testeissä mitattavan suureen arvo laskettiin näiden keskiarvona. Tulokset olivat huomattavasti paremmat kuin edellisillä malleilla[1], joten jatkokehitystä lähdettiin tekemään tältä pohjalta. Lisäetuna oli myös valitun menetelmän ohjaamattomuus.

[1] Valitettavasti tulosten tarkkaa numeerista arvoa ei ole enää saatavilla.

5.3 Ennusteen lähtötiedot ja tavoitetulos

Ennusteen lähtötietoina toimivat kullekin myyntipisteelle lasketut aiheryhmäprofiilit. Nämä profiilit on muodostettu laskemalla myyntipisteessä olevien lehtien viikkomyynnit aiheryhmittäin yhteen ja jakamalla lopputulos kokonaismyynnillä:

$$p_m(a) = \frac{\sum_{a(l)=a} v_m(l)}{\sum_l v_m(l)} \,. \tag{5.1}$$

Näin saatuja profiilivektoreita $\mathbf{p}_m = [p_m(a_1), p_m(a_2), \ldots, p_m(a_k)]$ käytettiin k-NN -menetelmän piirrevektoreina.

Maksimoitavana suureena toimi alustavasti lehden keskimääräinen viikkomyynti, ja uusi valikoima olisi näin ollen n parhaiten myyvää lehteä. Koska Lehtipisteen edustajat kokivat tämän kuitenkin diskriminoivan harvemmin ilmestyviä lehtiä, otettiin laskentaan lisäksi korjaustermiksi lehden keskimääräisestä viikkomyynnistä ja lehtien esiintymistiheydestä laskettu lehden keskimääräinen numeromyynti kappaleina:

$$n_m(l) = \frac{52}{N_l} \times v_m(l) \,. \tag{5.2}$$

Koko aineistosta laskettiin sekä $v_m(l)$:n että $n_m(l)$:n keskihajonnat. Koska $n_m(l)$:n keskihajonta oli jonkin verran pienempi, sen painoarvoa hieman kasvatettiin kertoimella $\lambda = 1{,}13$.

Näin ollen laskennassa käytetty myyntipistekohtainen hyvyysarvio $g_m(l)$ laskettiin seuraavasti:

$$g'_m(l) = v_m(l) + \lambda n_m(l) \tag{5.3}$$

$$g_m(l) = g'_m(l) \Big/ \sum_l g'_m(l) \,. \tag{5.4}$$

5.4 Ennusteiden laatiminen

k-NN -algoritmia laajennettiin hieman mallin luotettavuuden parantamiseksi. Koska myyntipisteiden naapurustoon saattoi päätyä myynniltään huomattavasti muusta joukosta poikkeavia myyntipisteitä, päätettiin näistä kahdestakymmenestä pudottaa pois sellaiset, joiden kokonaismyynti poikkesi 2,5 keskihajonnan verran keskiarvosta.

Lehtipisteen toimittamasta valikoimatiedostosta laskimme myynnissä olevien nimikkeiden määrän, joten kykenimme lisäksi arvioimaan myyntipisteen myynnin nimikettä kohden. Tätä käytettiin arvioimaan sitä, oliko myyntipiste myynniltään parempi kuin ennusteen kohteena oleva myyntipiste. Lopulliseen ennusteeseen otettiin mukaan vain ne myyntipisteet, jotka olivat todellakin parempia. Jos näitä ei löytynyt vähintään kolmea, valikoimaan ei ehdotettu muutoksia.

Tämän jälkeen osa myyntipisteen valikoimasta rajattiin ennusteiden ulkopuolelle siten, ettei valikoimaan ehdotettu näitä tuotteita uusina, muttei toisaalta myöskään ehdotettu niiden poistoa valikoimasta. Syyt tähän olivat pääasiassa liiketoiminnalliset, kuten tiettyjen lehtien hyvin rajattu ilmestymisalue (tiettyyn raviratan sidotuilla ravilehdillä) tai tiettyihin myyntipistetyyppeihin rajattu levitys.

On huomattava, että nämä tuotteet olivat myyntipisteiden samankaltaisuuksia tutkittaessa mukana; niiden rajaaminen sitä ennen olisi ollut informaation turhaa hävittämistä.

Kun parempien vertailumyyntipisteiden ja ennustettavien lehtien joukot olivat nyt selvillä, vertailumyyntipisteiden m lehdille l_m muodostettiin hyvyysarvio $g_m(l)$ kappaleen 5.3 mukaan. Tätä hyvyysarviota painotettiin lisäksi seuraavassa kuvatuin tavoin tiettyjen Lehtipisteen liiketoimintasääntöjen ja -tavotteiden saavuttamiseksi.

Näihin sääntöihin kuului muun muassa vertailumyyntipisteiden ruotsinkielisten lehtien hyvyysarvion skaalaaminen siten, että vertailumyyntipisteiden ja kohdemyyntipisteen ruotsinkielisten lehtien osuus myynnistä olisi sama. Samaten eräät aiheryhmät voivat olla myyntipisteen mielestä epäsoveliaita; näille aiheryhmiin kuuluvien lehtien hyvyysarvio asetettiin nollaan, mikäli myyntipisteestä ei löytynyt yhtään kyseisin aiheryhmiin kuuluvien lehtien myyntitapahtumia. Lisäksi myyntipisteen valikoimaan äskeittäin lisättyjä lehtiä haluttiin pitää valikoimassa myyntituloksista riippumatta lähtien siitä oletuksesta, että niiden todellisesta kysynnästä ei ole vielä tarpeeksi näyttöä.

Kohdemyyntipisteelle laskettiin vastaava ehdotushyvyysarvo $g^*(l)$ kaikkien vertailumyyntipisteiden m hyvyysarvioiden $g_m(l)$ keskiarvona. Kaavan 5.4 mukaisesti hyvyysarvion summa kussakin myyntipisteessä on 1, joten hyvyysarviota voi ajatella lehtiin liitettynä todennäköisyysjakaumana.

Lopullinen ennustettu $g_u(l)$ hyvyysarvio saatiin siirtämällä myyntipisteen vanhaa hyvyysarviota $g_v(l)$ hieman ehdotushyvyysarvion suuntaan ja lisäämällä mukaan koko aineistosta laskettu globaali hyvyysarvio $g_g(l)$:

$$g'_u(l) = \alpha g_v(l) + \beta g^*(l) + \gamma g_g(l) \tag{5.5}$$

$$g_u(l) = g'_u(l) \Big/ \sum_l g'_u(l) \ . \tag{5.6}$$

Kaavassa 5.5 eri termien painokertoimet valittiin Lehtipisteen kommenttien perusteella sopivaksi:

$$\alpha = 0{,}5, \ \beta = 1 \text{ ja } \gamma = 0{,}2. \tag{5.7}$$

Myyntipisteen uudeksi kokonaismyynniksi arvioitiin vertailumyyntipisteiden kokonaismyynnin keskiarvo.

Projektissa sovittiin Lehtipisteen kanssa, että valikoiman kokoon (n nimikettä) ei tehdä muutoksia. Tällöin ehdotus uudeksi valikoimaksi on n hyvyysarvion mukaan parasta nimikettä.

Luku 6

Tulokset

Mallin toimivuutta ja sen antamia ennusteita tutkittiin kolmessa vaiheessa. Ensiksi analysoimme mallin antamia tuloksia historiadatan puitteissa. Tämän jälkeen ennusteet annettiin Lehtipisteen asiantuntijoiden tutkittaviksi ja kritisoitaviksi. Lopulta mallin tuottamia ennusteita kokeiltiin kenttäolosuhteissa.

6.1 Mallin suorituskyvyn arviointi historiallisella aineistolla

Käytetty mallinnustekniikka ei varsinaisesti ota huomioon minkäänlaista testiaineistoa. Näin ollen emme voi arvioida mallin suoritus- ja yleistämiskykyä pelkästään tutkimalla sopivan virhefunktion käyttäytymistä testiaineistolla.

Koska jonkinlainen arvio mallin kyvyistä oli kuitenkin tarpeen, mutta mallin ennusteiden jatkuva analysointi asiantuntijajoukolla olisi hidastanut mallinnustyötä entisestään, kehitimme seuraavanlaisen menetelmän numeeristen tulosten saamiseen.

Kullekin myyntipisteelle voidaan muodostaa ehdotetuista muutoksista vektori $\mathbf{\Delta}_i$ siten, että vektorissa on kullekin lehdelle joko -1, 0 tai 1 sen mukaan mikäli lehti poistettiin valikoimasta, jätettiin valikoimaan tai lisättiin valikoimaan, vastaavasti. Samanlainen analyysi voidaan suorittaa myös kahden peräkkäisen myyntikauden tuloksille, tutkimalla mitkä lehdet olivat myynnissä kaudella T ja $T + 1$ ja laskemalla tästä aineistossa tapahtuneet muutokset.

Näiden kahden vektorin pistetulosta voidaan sitten todeta, onko tapahtunut muutos ollut samansuuntainen vai vastakkaissuuntainen tulosten kanssa. Vertaamalla tätä tulosta myyntipisteen myynnin suhteelliseen muutokseen voidaan myös katsoa, onko ennusteen ja todellisten tapahtumien samankaltaisuu-

della korrelaatiota myynnin paranemiseen vai ei.

Koska vertailtava syntyvä vektori pitää sisällään noin 2000 alkiota, ja historiadatassa tehdyille muutoksille on yleensä ollut useita vaihtoehtoja, oli odotettavissa että useimmissa tapauksissa muutokset eivät juurikaan osuneet kohdalleen. Näin myös kävi ja suurimmalla osalla pistetulon arvoksi tuli 0. Niillä myyntipisteillä, joilla pistetulo poikkesi nollasta, oli havaittavissa positiivinen korrelaatio lisämyynnin ja ennusteen samankaltaisuuden välillä[1].

6.2 Asiantuntijoiden analyysi

Kun malli toimi mielestämme tyydyttävällä tasolla, sen tuottamat ennusteet tarjottiin Lehtipisteen myyntiedustajille tutkittaviksi. Edustajien tuottaman kritiikin pohjalta mallin toimintaparametreihin tehtiin pieniä korjauksia mahdollisten epäkohtien parantamiseksi. Muutaman ylläkuvatun kaltaisen iteraatiokierroksen jälkeen Lehtipisteen edustajat olivatkin valmiit hyväksymään ennusteen kenttäkokeisiin.

Esimerkki Lehtipisteen edustajille toimitetusta ennusteesta on liitteessä A.

6.3 Kenttäkokeiden järjestelyt

Jotta mallin ennusteiden toimivuudesta saataisiin hyvä yleiskäsitys, mallia kokeiltiin kolmessa myyntipiirissä. Myyntipisteisiin sovellettujen rajausten jälkeen (katso kappale 4.1.7, s. 28) käytössä oli noin 500 myyntipistettä eri puolilta Suomea. Nämä myyntipisteet yhdistettiin pareiksi samankaltaisuuden mukaan. Toinen parin myyntipisteistä valittiin pilottipisteeksi ja sille laskettiin uusi valikoima, kun taas jäljellejäävä myyntipiste määrättiin verrokkipisteeksi, jonka valikoimaan ei tehty muutoksia.

Nämä vastinparit etsittiin lähtemällä liikkelle aikaisemmin tuotetusta myyntipisteiden klusteroinnista. Parit valittiin laskemalla piirrevektoreiden väliset etäisyydet ja muodostamalla vastinpari niistä vielä parittamattomista myyntipisteistä, joiden välinen etäisyys oli kaikkien pienin. Näin saatiin noin 250 paria.

Parien määrää kuitenkin karsittiin tulosten arvioinnin aiheuttaman työmäärän vuoksi. Loppujen lopuksi testeissä käytettäväksi määräksi sovittiin 25 paria/piiri. Parit poimittiin jo muodostetuista pareista ajan säästämiseksi.

[1] Valitettavasti tarkkaa numeerista arvoa ei ole enää saatavilla.

Edustajat valitsivat näistä 75 parista, kumpi pisteistä otetaan verrokkipisteeksi ja kumpi pilottipisteeksi. Mikäli vastinparit olivat luonteeltaan rajusti erilaiset, tai oli syytä odottaa, että kumpikaan ei suostuisi kokeeseen, edustajat saattoivat myös rajata myyntipisteen kokonaan pois. Näin ollen testattavien myyntipisteiden joukkoon jäi 64 myyntipistettä.

Kenttäkokeissa vaihdettiin myyntipisteen valikoima mallin ehdottamaksi ja verrattiin myyntiä ennen ja jälkeen muutoksen. Testikausi aloitettiin marraskuun alussa, ja se kesti kolme kuukautta päättyen helmikuun alussa. Vastaavaksi vertailukaudeksi valittiin syys- ja lokakuu, jotta kesäsesongin vaikutus ei vääristäisi tuloksia.

Testikauden aikana neljä myyntipistettä lopetti toimintansa tai laajensi sitä niin huomattavasti, että ne ja niiden vastinparit oli jätettävä tarkastelun ulkopuolelle. Lopulliseksi testattavaksi määräksi jäi siis 60 pilottipistettä, joita vastasi 60 verrokkipistettä.

Tulosten vertailussa piti ottaa seuraavia seikkoja huomioon:

- vertailukausi ja testikausi olivat eripituiset,
- lehtien myynnissäolojaksojen pituudet vaihtelivat, joten osalla lehdistä myyntijakso saattoi jäädä kesken,
- kausivaihtelun merkitystä eri kuukausina ei tunnettu ja
- vastinparien myynti saattoi poiketa toisistaan.

Näiden tekijöiden pohjalta tutkittavaksi suureeksi valittiin testikauden ja vertailukauden myyntien suhde.

6.4 Tilastollisen testauksen määrittely

Tutkittavina näytejoukkoina ovat ylläkuvatun mukaisesti vertailu- ja testikauden myyntien suhteet sekä verrokkipisteistä että pilottipisteistä. Vertailussa päätettiin jättää vastinpaririippuvuudet huomiotta, sillä verrokki- ja pilottipisteet liitettiin aikoinaan pareiksi samanlaisten myyntipisteiden löytämistä varten, eikä niillä ole tilastollisessa analyysissä tarkoitettua riippuvuutta.

Näin ollen tutkittavana on kaksi datajoukkoa, joiden varianssit ovat tuntemattomia eikä niitä voida olettaa samoiksi. Datajoukot oletetaan normaalijakautuneiksi, mutta oletusta on vaikea testata, koska käytettävissä olevaa dataa on liian vähän asianmukaisten testien suorittamiseen.

Testauksessa halutaan tietää, onko pilottijoukon keskiarvo eri kuin verrokkijoukon. Meitä kiinnostaa erityisesti myönteinen tulos, joka indikoi mallin kasvattaneen myyntiä eri myyntipisteissä. Myös vastakkaisen tuloksen mahdollisuuteen on varauduttava.

Nollahypoteesi ja vaihtoehtoinen hypoteesi ovat tällöin:

$$H_0 : \mu_p = \mu_v, \qquad H_1 : \mu_p \neq \mu_v, \tag{6.1}$$

jossa μ_p ja μ_v on pilottijoukon ja verrokkijoukon keskiarvot.

Tällaisessa tilanteessa soveltuva tilastollinen testi on kaksisuuntainen t-testi, jonka testisuure määritellään Lainisen [22] mukaan yleisessä tapauksessa seuraavasti:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}, \tag{6.2}$$

Ylläolevassa N_i ovat joukon i näytteiden lukumäärä, $\bar{Y}_i$ ovat vastaavat näytekeskiarvot ja s_i^2 vastaavat näytevarianssit. t-jakauman vapausasteiden määrä ν saadaan kaavasta

$$\nu = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_1^2/N_2)^2/(N_2 - 1)}\,. \tag{6.3}$$

6.5 Kenttäkokeiden tulokset

Tulokset koko aineistolle esitetään taulukossa 6.1, piireittäin jaoteltuna taulukossa 6.2 ja toimialoittain jaoteltuna taulukossa 6.3. Taulukon sarakkeina ovat datapisteiden määrä, pilottipisteiden parannus verrattuna verrokkipisteisiin, vapausasteiden määrä ν, t-testisuureen arvo ja t-jakaumasta määritetty P-arvo.

Koska koemyyntipisteiden määrää vähennettiin rajusti, tulosten tilastollinen luotettavuus useimmissa tapauksissa kärsii. Esimerkiksi kahdesta toimialasta on aineistossa (katso taulukko 6.3) on aineistossa vain yksi esimerkki; näille tapauksille ei ole laskettu t-testisuuretta tai P-arvoa.

Tilastollisen analyysin mielessä tulosten tilastollinen merkitsevyys on huono. Traditionaalisella 5 % riskitasolla vain huoltoasemat paransivat tulostaan tilastollisesti merkittävästi. 10 % riskitasolla olisi myös R-kioskit ja pienmyymälät voitu laskea onnistuneiden joukkoon. Vastaavasti myös hypermarketit ovat 10 % riskitason tuntumassa, mutta valitettavasti niiden tulos heikkeni testin aikana.

N	Muutos	ν	t	P
60	2,36%	114	0,988	0,325

Taulukko 6.1: t-testin tulokset koko aineistolle

Piiri	Muutos	N	ν	t	P
2 (Itäinen Uusimaa)	-1,13%	18	33	-0,314	0,756
26 (Länsirannikko)	5,15%	21	39	1,165	0,251
30 (Pohjois-Suomi)	2,25%	21	34	0,639	0,527

Taulukko 6.2: t-testin tulokset jaoteltuna myyntipiireittäin

Toimiala	Muutos	N	ν	t	P
Pienmyymälä	29,88%	1	—	—	—
Kahvila, ravintola	20,75%	2	2	0,083	0,941
Huoltoasema, huoltoasema-kahvila	13,10%	8	10	2,377	0,038
Kauppiaskioski	13,02%	2	2	-0,034	0,976
Supermarket, pieni	11,87%	6	8	0,587	0,574
R-kioski	5,37%	9	15	-1,762	0,098
Kirja- ja paperikauppa	9,59%	1	—	—	—
Iso valintamyymälä	-0,04%	7	10	-0,337	0,743
Kioski, kanttiini	-3,42%	6	10	-0,156	0,879
Hypermarket	-5,61%	5	7	1,673	0,138
Pieni valintamyymälä	-5,85%	4	4	1,360	0,245
Supermarket, iso	-8,62%	9	13	1,504	0,157

Taulukko 6.3: t-testin tulokset jaoteltuna toimialoittain

Luku 7

Yhteenveto ja johtopäätökset

Tässä kappaleessa käydään läpi mallin menestystä kenttäkokeissa ja pohditaan sen heikkouksia ja vahvuuksia. Lisäksi kappaleessa käsitellään muutamia mahdollisia jatkokehityskohteita.

7.1 Mallin toiminta

Mallin tulokset olivat lupaavia; vaikka täydellinen tilastollinen varmuus asiasta jäi saavuttamatta, malli kuitenkin suuressa osassa tapauksista tuotti positiivisen muutoksen. Lehtipisteen myynti vuonna 2003 oli 247,1 miljoonaa euroa [28], joten kokonaisuudelle projisoituna 2,36 % parannus on 5,8 miljoonaa euroa vuodessa.

Sikäli kun tuloksia luonnehtii tilastollisesta merkitsevyydestä huolimatta, on nähtävissä jonkinlainen käänteinen riippuvuus myyntipisteen koon ja testituloksen hyvyyden välillä. Pienten myyntipisteiden hyvä testimenestys on intuitiivisesti selitettävissä. Ensinnäkin, pieniä myyntipisteitä on paljon, ja myyntiedustajan aika ei aina riitä huolelliseen analyysiin, vaan aika ja tarmo on säästettävä isompia ja tärkeämpiä asiakkaita kohtaan. Toisekseen ehdotettujen korjaustoimenpiteiden ei tarvitse olla mitenkään monimutkaisia; pienessä myyntipisteessä muutaman lehden vaikutus voi olla huomattava.

Esimerkkinä voidaan ajatella pieniä huoltamoita. Koska huoltamoiden valikoimat ovat tyypillisesti pieniä, 20–40 lehteä, ja mahdollisia valikoimiin kelpaavia lehtiä on noin 1500, valikoimissa ei välttämättä ole montakaan yhteistä lehteä. Ihmisen olisi tutkittava usean huoltamon valikoimaa ja myyntihistoriaa muodostaakseen käsityksen parhaista lehdistä, kun taas automaattinen järjestelmä voi helposti tunnistaa heikot lehdet ja ehdottaa korvaavia.

Suurten myyntipisteiden heikkoa menestystä ei voi suoralta kädeltä analysoida yhtä tyhjentävästi. Kyseisissä myyntipisteissä on yleensä varsin laaja valikoima, ja näin ollen niihin on vaikea ehdottaa enää uusia hyvin menestyviä lehtiä, joten parannusten vaikutus jää pienemmäksi.

Toisaalta Lehtipiste kertoo vuosikatsauksessaan [28] lehtimyynnin kasvaneen voimakkaimmin S-ryhmän myyntipisteissä. Kenttäkokeissa heikoiten menestyneiden hypermarkettien ja isojen supermarkettien verrokkipisteet olivat enimmäkseen S-ryhmän myyntipisteitä. Sen sijaan pilottipisteet kuuluivat enimmäkseen K-ryhmään. Yllä oleva huomioon ottaen tämä voi olla hyvinkin ratkaisevaa, etenkin valitun testaustekniikan puitteissa.

Näin ollen tilanne vaatii vielä asiantuntija-analyysia muiden vastaavien koejärjestelyjen ulkopuolisten tekijöiden selvittämiseksi.

Pienellä jatkokehityksellä mallista voi hyvinkin kehkeytyä eri toimialoille sovellettava tehokas ratkaisu.

7.2 Jatkokehitys

Mallin jatkokehitys voidaan jakaa kahteen luokkaan: nykymenetelmään tehtävät parannukset, jotka eivät ratkaisevasti muuta mallin luonnetta, ja kokonaan uudet mallinnusmenetelmät, jotka käyttävät tässä projektissa selvitettyjä seikkoja hyödykseen.

7.2.1 Nykyisen mallin parannukset

Malli huomioi ruotsinkielisten myyntipisteiden erityistarpeet keskimääräisesti hyvin, mutta yksittäisten myyntipisteiden kohdalla virheet voivat olla huomattavia. Näin ollen kielisyys on otettava tarkempaan käsittelyyn.

Kuten taulukosta B.2 on laskettavissa, neljä yleisintä kieltä kattavat 94,4% kaikista lehdistä. Kriittistä on kuitenkin suomen ja ruotsin suhde, joten eräs tutkimisen arvoinen vaihtoehto voisi olla hienovaraisempi aiheryhmäprofiili, jossa kustakin aiheryhmästä laskettaisiin erikseen suomen-, ruotsin- ja muunkielisten lehtien osuudet.

Nykymallin suorituskykyä voisi myös kehittää tutkimalla vaihtoehtoisia samankaltaisuuden määritelmiä. Esimerkiksi Kullback-Leibler -divergenssin soveltaminen mallin k-NN -vaiheessa voisi myös olla hedelmällistä. Kyseisessä vaiheessa tutkitaan myyntipisteiden samankaltaisuutta yhden tietyn myyntipisteen näkökulmasta, jolloin KL-divergenssin matemaattiset rajoitteet eivät ole ongelmallisia.

Lisäksi nykyratkaisu ehdotti vain korvauksia, se ei omaehtoisesti ehdottanut lehtien määrää lisättäväksi. Tämä johtui siitä, ettei luotettavaa arvioita myyntipisteiden hyllytilan määrästä ole saatavilla. Tällainen arvio tai tieto mahdollistaisi myös vajaiden valikoimien täydentämisen.

7.2.2 Vaihtoehtoiset mallinnusratkaisut

Vaihtoehtoisia mallinnusmenetelmiä tutkittaessa kannattaisi lähteä liikkeelle hyvyysarvion matemaattisesta analyysistä. Kyseessä on periaatteessa yksinkertainen prosessi ja sen tulokset ovat helposti tulkittavissa, mutta sen matemaattisia implikaatioita ja ongelmia ei ole kuitenkaan kattavasti analysoitu.

Lisäksi malliin tehtiin Lehtipisteen pyynnöstä säätöjä ja korjauksia. Osa näistä korjauksista on matemaattisesti perustelemattomia; mallista löytyy erinäisiä vakioita, joiden arvoa muuteltiin, kunnes Lehtipisteen edustajat olivat tyytyväisiä tulokseen. Näin ollen mallin tarkempi matemaattinen analyysi olisi tarpeen; etenkin mallia voisi tulkita bayesiläisen todennäköisyyslaskennan valossa ja koettaa tunnistaa mahdollisia epäoptimaalisuuksia tältä kantilta.

Mikäli hyvyysarvio on todellakin tulkittavissa todennäköisyysjakaumaksi, tätä seikkaa voisi hyödyntää enemmän ennusteiden laatimisessa. Esimerkiksi uuden valikoiman (tai tietyn osuuden siitä) voisi valita näytteistämällä tätä jakaumaa; nykyinen valikoimanmuodostustapa on mahdollisesti epäreilu pienilevikkisille lehdille ja on kuviteltavissa tapaus, jolloin valikoima konvergoituisi koko maassa muutamaan sataan yleisimpään lehteen. Samaten algoritmia voisi tällöin pitää eräänlaisena geneettisenä algoritmina, jossa k-NN olisi risteytysvaihe ja näytteistys mutaatiovaihe.

Hyvyysarvion probabilistisesta tulkinnasta voi myös edetä probabilistiseen mallintamiseen. Kuten kappaleessa 2.5 mainittiin, tutkimusongelman relevanteille osille on esitettävissä yhteistapahtumadatatulkinta. Tällöin jokin kappaleessa mainituista mikstuuri- tai komponenttimalleista soveltuisi mainiosti ongelmaan.

Liite A

Esimerkkiennuste

Ennustettava myyntipiste on piirissä 30 (Pohjois-Suomessa) toimiva pieni supermarket. Taulukossa A.1 esitellään lehdet, joita malli ehdotti poistettavaksi. Lehdet on järjestetty ennustetun myynnin mukaan. Jokaisesta lehdestä ilmoitetaan aiheryhmä, perusmäärä (myyntipisteeseen toimitettujen lehtien määrä), alkuperäisestä aineistosta laskettu viikkomyynti euroina, sekä ennusteen määrittämänä viikkomyynti euroina, kappalemäärä viikossa ja kappalemäärä numeroa kohti.

Taulukossa A.2 on vastaavasti tilalle ehdotettujen lehtien lista. Taulukon kenttien merkitys on lähes sama kuin taulukossa A.1, seuraavin eroin:

- Ehdotettu perusmäärä on laskettu Lehtipisteen ehdottamalla kaavalla $\lceil k \rceil + \lfloor k/10 \rfloor + 1$, jossa k on mallin ennustama numerokohtainen kappalemyynti.
- Lehti on saattanut olla myyntipisteessä aiemmin, jolloin alkuperäinen myynti poikkeaa nollasta.

Aiheryhmä	Perusmäärä	Myynti	Ennuste	Kpl/vk	Kpl/nro
Asuminen	3	0,00 e	0,37 e	0,1	1,6
Asuminen	2	0,00 e	0,37 e	0,1	1,7
Ristikot	2	0,23 e	0,38 e	0,1	1,6
Perhelehdet	3	0,19 e	0,40 e	0,1	0,4
Ristikot	2	0,63 e	0,44 e	0,2	1,3
Infotekniikka	5	0,50 e	0,48 e	0,1	0,5
Infotekniikka	2	0,44 e	0,53 e	0,1	0,7
Harrastukset	3	0,67 e	0,55 e	0,1	0,8
Naistenlehdet	2	0,88 e	0,58 e	0,1	0,9
Terveys, kauneus ja kuntoilu	2	0,00 e	0,58 e	0,1	0,3
Ilmoituslehdet	1	0,52 e	0,58 e	0,2	1,6
Käsityölehdet	2	0,50 e	0,65 e	0,1	1,1
Harrastukset	5	0,45 e	0,66 e	0,1	1,8
Sarjakuva-albumit	5	0,60 e	0,73 e	0,2	2,0
Ilmoituslehdet	1	0,41 e	0,94 e	0,3	1,2
Naisten lukemistot	1	0,98 e	0,96 e	0,2	1,0
Ilmoituslehdet	3	0,99 e	0,96 e	0,3	0,7
Nuorisolehdet	2	0,85 e	1,27 e	0,3	1,3

Taulukko A.1: Mallin ehdottamat poistot

Aiheryhmä	Perusmäärä	Myynti	Ennuste	Kpl/vk	Kpl/nro
Sarjakuvalehdet	3	1,52 e	1,07 e	0,4	1,8
Ristikot	4	0,00 e	1,07 e	0,3	2,0
Ristikot	9	0,00 e	1,13 e	0,3	7,1
Ristikot	4	0,00 e	1,27 e	0,6	2,7
Ristikot	4	0,00 e	1,29 e	0,5	2,0
Sarjakuvalehdet	2	0,92 e	1,33 e	0,3	1,4
Ristikot	5	0,00 e	1,55 e	0,3	4,5
Käsityölehdet	3	0,00 e	1,66 e	0,2	1,1
Infotekniikka	3	3,07 e	1,73 e	0,3	1,2
Sarjakuvalehdet	4	2,37 e	1,74 e	0,5	2,1
Sarjakuvalehdet	3	2,85 e	1,89 e	1,0	1,0
Sarjakuvalehdet	3	1,89 e	2,00 e	0,6	1,2
Asuminen	3	0,94 e	2,16 e	0,3	1,5
Sarjakuvalehdet	4	3,26 e	2,46 e	0,4	2,7
Asuminen	3	0,00 e	2,63 e	0,4	1,9
Infotekniikka	4	4,08 e	3,42 e	0,6	2,7
Naistenlehdet	4	0,00 e	3,63 e	0,6	2,8
Asuminen	4	4,78 e	3,67 e	0,5	2,2
Autolehdet	5	6,50 e	4,88 e	0,8	3,0

Taulukko A.2: Mallin ehdottamat lisäykset

Liite B

Kategorisia muuttujia

Taulukossa B.1 esitetään aineistoon tehtyjen rajauksien jälkeen lehtien lukumäärät julkaisumaiden mukaan suurimmasta pienimpään. Tehdyt rajaukset on kuvattu kappaleessa 4. Vastaavasti taulukoissa B.2 ja B.3 esitetään lehtien kielet ja aiheryhmät.

Myyntipisteistä esitellään vastaavasti taulukossa B.4 toimialajakauma.

Maa	N	Maa	N
Suomi	532	Eesti	8
Englanti	315	Venäjä	6
Usa	249	Tanska	2
Saksa	171	Sveitsi	2
Ruotsi	131	Norja	1
Ranska	37	Alankomaat	1
Espanja	27	Tsekin tasavalta	1
Italia	23	Australia	1

Taulukko B.1: Julkaisumaajakauma

Kieli	N	Kieli	N
Englanti	603	Espanja	25
Suomi	502	Italia	16
Ruotsi	159	Venäjä	8
Saksa	150	Eesti	7
Ranska	37		

Taulukko B.2: Kielijakauma

Aiheryhmä	N	Aiheryhmä	N
Asuminen	126	Perhelehdet	44
Harrastukset	111	Lasten- ja nuortenkirjat	40
Autolehdet	99	Trendilehdet	37
Infotekniikka	91	Ruoka ja juhlat	32
Erotiikka	89	TV- ja kulttuurilehdet	30
Naistenlehdet	88	Sarjakuva-albumit	29
Urheilu	86	Ilmoituslehdet	27
Käsityölehdet	71	Rahapeli- ja ravilehdet	24
Uutis-,talous- ja tiedelehdet	69	Taskukirjat	20
Ristikot	61	Sarjakirjat	18
Musiikki	60	Venelehdet	16
Moottorilehdet	59	Muut kirjat	8
Terveys, kauneus ja kuntoilu	54	Naisten lukemistot	5
Sarjakuvalehdet	54	Miesten lukemistot	3
Nuorisolehdet	45	Puuhalehdet	2

Taulukko B.3: Aiheryhmäjakauma

Toimiala	N
Iso valintamyymälä	819
Pieni valintamyymälä	644
Supermarket, pieni	604
Kioski, kanttiini	539
Huoltoasema, huoltoasema-kahvila	488
R-Kioski	469
Supermarket, iso	364
Kauppiaskioski	256
Erikoisliikkeet	205
Hypermarket	126
Tavaratalo	87
Kirja- ja paperikauppa	86
Kahvila, ravintola	83
Pienmyymälä	79
Ei toimialaa	67
Näytelehtipiste (LP-linja)	35
Hotelli, motelli, matkustajakoti	30
Sesonkimyymälät	14
Lähettämöt	13
Veikkausrasti	1

Taulukko B.4: Toimialajakauma

Kirjallisuutta

[1] Karlos A. Artto. *A Supply Decision Procedure for the Distribution of Magazines.* Number 71 in Acta Polytechnica Scandinavica, Mathematics and Computing in Engineering Series. Finnish Academy of Technology, 1994.

[2] Albert-László Barabási. *Linked — The New Science of Networks.* Perseus Publishing, Cambridge, MA, USA, April 2002.

[3] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270, 1994.

[4] Michael Berthold and David J. Hand, editors. *Intelligent Data Analysis.* Springer-Verlag, 1999.

[5] Ella Bingham, Heikki Mannila, and Jouni K. Seppänen. Topics in 0-1 data. In David Hand, Daniel Keim, and Raymond Ng, editors, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pages 450–455, Edmonton, Alberta, Canada, July 2002.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS*14).* MIT Press, 2002.

[7] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting And Control.* Holden-Day, San Francisco, 1970.

[8] W. Buntine. Variational extensions to EM and multinomial PCA. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning, Helsinki, Finland, August 19–23, 2002. Proceedings*, pages 23–34. Springer-Verlag, 2002.

[9] W. Buntine and S. Perttu. Is multinomial PCA multi-faceted clustering or dimensionality reduction. In C.M.Bishop and B.J. Frey, editors, *Proc. 9th Int. Workshop on Artificial Intelligence and Statistics*, pages 300–307. Society for Artificial Intelligence and Statistics, 2003.

[10] Chris Chatfield. *Time-series forecasting.* Chapman & Hall/CRC, 2000.

[11] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[13] Simon Haykin. *Neural networks: a comprehensive foundation.* Prentice-Hall, second edition, 1999.

[14] Frederick S. Hillier and Gerald J. Lieberman. *Introduction to Operations Research.* McGraw-Hill, eigth edition, 2005.

[15] Thomas Hofmann. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference in Artificial Intelligence (IJCAI'99).* Morgan Kaufmann.

[16] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[17] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[18] Thomas Hofmann and Jan Puzicha. Statistical models for co-occurrence data. A.I. Memo 1625, Artificial Intelligence Laboratory, MIT, February 1998.

[19] Thomas Hofmann, Jan Puzicha, and Michael I. Jordan. Learning from dyadic data. In *Advances in Neural Information Processing Systems (NIPS*11)*, pages 466–472. Neural Information Processing Systems Foundation, MIT Press, 1999.

[20] Jan Holmström. Handling product range complexity — a case study on re-engineering demand forecasting. *Business Process Management*, 4(3):241–258, 1998.

[21] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, New York, third edition, 2001.

[22] Pertti Laininen. Sovelletun todennäköisyyden kaavoja ja taulukoita. Edita Opetusmoniste, 2000.

[23] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[24] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Processing Systems (NIPS*13)*, pages 556–562. MIT Press, 2001.

[25] Michael Levy and Barton A. Weitz. *Retail Management*. Irwin/McGraw-Hill, third edition, 1998.

[26] Kaija Pöysti. Aikakauslehtien irtomyynnin ennustemalli. Diplomityö, Teknillinen korkeakoulu, 1985.

[27] Lasse Rasinen. HACM-mallin evaluointi. T-61.195 Informaatiotekniikan erikoistyö I.

[28] Rautakirja Oy. *Vuosikatsaus 2003*, page 11. 2004.

[29] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. Self-Organizing Map in Matlab: the SOM Toolbox. In *Proc. of Matlab DSP Conference 1999*, pages 35–40, Espoo, Finland, 1999.

[30] Juha Vesanto. *Data Exploration Process Based on the Self-Organizing Map*. Number 115 in Acta Polytechnica Scandinavica, Mathematics and Computing Series. Finnish Academy of Technology, Espoo, Finland, 2002. ISBN 951-666-596-9, ISSN 1456-9418.

[31] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inform Theory*, 37(4):1085–1094, July 1991.

[32] Xtract Ltd. *Consumer LifeCycles White Paper*, August 2004. Saatavilla pyynnöstä.